Ensemble forecasting using sequential aggregation for photovoltaic power applications

Jean Thorey

▶ **To cite this version:**

Jean Thorey. Ensemble forecasting using sequential aggregation for photovoltaic power applications. Statistics [math.ST]. Université Pierre et Marie Curie - Paris VI, 2017. English. NNT : 2017PA066526 . tel-01697133v2

HAL Id: tel-01697133
https://inria.hal.science/tel-01697133v2

Submitted on 27 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse de doctorat de l'Université Pierre et Marie Curie

École doctorale de Sciences Mathématiques de Paris Centre
Laboratoire : Équipe-projet INRIA CLIME & CEREA.

par

Jean Thorey

Prévision d'ensemble par agrégation séquentielle appliquée à la prévision de production d'énergie photovoltaïque.

Directrice de thèse :	Mme. Isabelle Herlin	Directeur de recherche (INRIA)
Encadrant de thèse :	M. Vivien Mallet	Chargé de recherche (INRIA)
Composition du jury :		
Rapporteurs :	Mme. Liliane Bel	Professeur (AgroParisTech)
	M. Jochen Broecker	Professeur (Université de Reading)
Examinateurs :	Mme. Petra Friederichs	Professeur (Université de Bonn)
	M. Olivier Mestre	Chercheur (Météo-France)
	M. Olivier Wintenberger	Professeur (Université Pierre et Marie Curie)
Encadrant industriel :	M. Christophe Chaussin	Ingénieur de recherche (EDF R&D)

Prévision d'ensemble par agrégation séquentielle appliquée à la prévision de production d'énergie photovoltaïque.

Résumé : Notre principal objectif est d'améliorer la qualité des prévisions de production d'énergie photovoltaïque (PV). Ces prévisions sont imparfaites à cause des incertitudes météorologiques et de l'imprécision des modèles statistiques convertissant les prévisions météorologiques en prévisions de production d'énergie. Grâce à une ou plusieurs prévisions météorologiques, nous générons de multiples prévisions de production PV et nous construisons une combinaison linéaire de ces prévisions de production. La minimisation du Continuous Ranked Probability Score (CRPS) permet de calibrer statistiquement la combinaison de ces prévisions, et délivre une prévision probabiliste sous la forme d'une fonction de répartition empirique pondérée. Dans ce contexte, nous proposons une étude du biais du CRPS et une étude des propriétés des scores propres pouvant se décomposer en somme de scores pondérés par seuil ou en somme de scores pondérés par quantile. Des techniques d'apprentissage séquentiel sont mises en oeuvre pour réaliser cette minimisation. Ces techniques fournissent des garanties théoriques de robustesse en termes de qualité de prévision, sous des hypothèses minimes. Ces méthodes sont appliquées à la prévision d'ensoleillement et à la prévision de production PV, fondée sur des prévisions météorologiques à haute résolution et sur des ensembles de prévisions classiques.

Ensemble forecasting using sequential aggregation for photovoltaic power applications.

Abstract: Our main objective is to improve the quality of photovoltaic power forecasts deriving from weather forecasts. Such forecasts are imperfect due to meteorological uncertainties and statistical modeling inaccuracies in the conversion of weather forecasts to power forecasts. First we gather several weather forecasts, secondly we generate multiple photovoltaic power forecasts, and finally we build linear combinations of the power forecasts. The minimization of the Continuous Ranked Probability Score (CRPS) allows to statistically calibrate the combination of these forecasts, and provides probabilistic forecasts under the form of a weighted empirical distribution function. We investigate the CRPS bias in this context and several properties of scoring rules which can be seen as a sum of quantile-weighted losses or a sum of threshold-weighted losses. The minimization procedure is achieved with online learning techniques. Such techniques come with theoretical guarantees of robustness on the predictive power of the combination of the forecasts. Essentially no assumptions are needed for the theoretical guarantees to hold. The proposed methods are applied to the forecast of solar radiation using satellite data, and the forecast of photovoltaic power based on high-resolution weather forecasts and standard ensembles of forecasts.

Remerciements

Je tiens à remercier tout d'abord mon superviseur de thèse Vivien Mallet, qui m'a fait confiance au cours de ces trois années en me proposant un sujet de thèse riche et une grande liberté pour avancer. Merci à toi Vivien pour ta grande disponibilité, ta patience et ton sens aigü du travail bien fait qui m'ont permis de progresser énormément.

Mes remerciements vont également à Jochen Broecker et Liliane Bel pour avoir accepté de rapporter ma thèse, ainsi qu'à Petra Friederichs, Olivier Mestre et Olivier Wintenberger en tant que membres du jury. Tout au long de ce travail, le comité de suivi de thèse composé notamment de Laurent Dubus, Luc Musson-Genon, Laurent Descamps, Yannig Goude et Philippe Blanc, m'a accompagné avec leurs précieux conseils. Merci également à Gilles Stoltz et Pierre Gaillard pour avoir répondu à mes questions théoriques.

Ces trois années ont passé à la fois très vite et très lentement, elles m'ont permis de rencontrer de nombreuses personnes comme Isabelle Herlin et tous les collègues de l'INRIA : Nathalie, Yann, Paul, Sylvain, Nicolas, Raphaël, Julien et Guillaume. Grâce à toi Isabelle, j'ai pu découvrir un peu les rouages d'une institution aussi riche que l'INRIA. Merci à tous pour votre enthousiasme et votre brin de folie, vos avis tranchés lors de débats passionnés autour du billard ou d'un tour de magie. Côté EDF, je souhaite remercier les groupes Météo appliquée et Environnement atmosphérique et plus particulièrement Stéphanie et Christophe pour leur gentillesse et leur attention, ainsi qu'Alain, l'homme au short qui ne sèche jamais, mais aussi Bénédicte, Sylvie, Augustin et tous les autres.

Je tiens bien sûr à remercier ma famille pour son soutien indéfectible, les amis de toujours et les nouvelles personnes qui sont entrées dans ma vie. Et plus particulièrement ma femme Camille qui est toujours à mes côtés.

Contents

État de l'art et contributions

L'amélioration des prévisions de production d'énergie photovoltaïque (PV) contribue à une meilleure intégration de l'énergie photovoltaïque. Les prévisions météorologiques, les modèles de conversion météo-production et les techniques de post-traitement statistiques constituent trois axes d'amélioration. Pour notre part, le problème est abordé de la façon suivante : un prévisionniste, souhaitant fournir des prévisions probabilistes de production PV, récupère des prévisions météorologiques (potentiellement de sources variées). Dans ce cadre général, de multiples méthodes peuvent être testées et combinées. Les prévisions météorologiques constituent la base de nos méthodes de post-traitement qui impliquent des ensembles de prévisions. Un état de l'art des modèles statistiques utilisés pour le PV est proposé par Bacher et al. [BMN09] et Inman et al. [IPC13]. La récente revue de l'état de l'art [Ant+16] inclut une revue des techniques de prévision probabiliste et de prévision d'ensemble appliquées pour le PV. Un nombre restreint de publications mentionnent des techniques de post-traitement prenant en compte un ensemble de prévisions météorologiques pour le PV [Zam+14 ; Ale+15 ; SAM16]. Par ailleurs, Lorenz et al. [LKH12] combinent des prévisions météorologiques et des prévisions issues de données satellitaires.

Un prévisionniste disposant de multiples prévisions peut souhaiter les combiner de façon optimale, par exemple par agrégation séquentielle. Le prévisionniste donne alors un poids à chaque prévision et fournit la combinaison linéaire des prévisions. Ces poids sont déterminés par une règle de mise à jour qui ne prend en compte que l'information du passé disponible à chaque pas de temps. De plus, la prévision fournie dispose de garanties théoriques de performance valables pratiquement sans hypothèses (sur un a priori, un processus stochastique ou une distribution sous-jacente), cf. Cesa-Bianchi et Lugosi [CL06], la revue de Shalev-Shwartz [Sha11] ou l'introduction de Stoltz [Sto10] en français. Les garanties théoriques sont essentiellement des garanties théoriques de robustesse (borne de regret), assurant que la combinaison des prévisions est sur le long terme au moins aussi performante que la meilleure prévision ou la meilleure combinaison de prévisions à poids fixes. Ces techniques ont déjà été testées sur des jeux de données variés : consommation d'électricité, concentration d'ozone, champs de vent et de pression [Sto10 ; MSM09 ; Mal10 ; Bau15 ; GGN16].

Nous montrons que l'agrégation séquentielle permet d'améliorer les prévisions d'ensoleillement au chapitre 2, d'après Thorey et al. [Tho+15]. Dans ce travail, nous étudions les ensembles de prévisions de 6 centres de prévision météorologique : China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), MetOffice (UKMO), Korea Meteorological Administration (KMA), Centro de Previsao Tempo e Estudos Climaticos (CPTEC), et Météo France (M.-F.), pour un total de 158 prévisions. Ces prévisions sont comparées aux observations d'origine

satellitaire de la base de données HélioClim-3, sous forme d'un champ d'observations sur la France métropolitaine. Les observations satellitaires de notre jeu de données ont une résolution de $1/12°$, qui est plus fine que la résolution de $0.25°$ à laquelle nous récupérons les prévisions. Nous trouvons que les ensembles de prévisions sont généralement sous-dispersés. Nous utilisons l'agrégation séquentielle pour combiner linéairement les 158 prévisions avec des poids pouvant varier en temps et en espace. À chaque point de grille i et pour chaque pas de temps t, nous utilisons une régression ridge escomptée pour trouver les poids $\boldsymbol{u}_t^{(i)}$ minimisant une certaine distance entre les prévisions $\boldsymbol{x}_{m,t'}^{(i)}$ et les observations $y_{t'}^{(i)}$ passées. Les poids sont choisis de sorte qu'ils minimisent

$$J(\boldsymbol{u}) = \lambda \|\boldsymbol{u} - \boldsymbol{w}_{\text{ref}}\|_2^2 + \sum_{t'=1}^{t-1} \left(1 + \frac{\gamma}{(t-t')^2}\right) (y_{t'}^{(i)} - \boldsymbol{u} \cdot \boldsymbol{x}_{t'}^{(i)})^2 \,,$$

où $\boldsymbol{w}_{\text{ref}}$ est un vecteur de réference, λ est le paramètre de régularisation et γ est le paramètre d'escompte. Cette méthode diminue l'erreur de prévision de 20% par rapport à la prévision de référence HRES du ECMWF. De plus, cette méthode corrige les biais locaux spatiaux et produit des motifs spatiaux plus réalistes de l'ensoleillement.

Dans la littérature météorologique, la minimisation du CRPS est une stratégie commune pour calibrer les prévisions probabilistes [Gne+05 ; Sch14 ; JMA15]. Cependant, les techniques classiques n'offrent pas de garantie théorique de robustesse et ont recours à des hypothèses sur les distributions. Par exemple, le « Bayesian model averaging » (BMA) fournit un mélange de distributions paramétriques, une somme de gaussiennes [Gne+05] ou de distributions gamma [SGR10 ; Slo+07]. La régression non-homogène cale les paramètres d'une distribution paramétrique en fonction des attributs d'un ensemble de prévisions [Gne+05 ; Wil09 ; TG10]. Ainsi la moyenne et la variance d'une distribution gaussienne sont calées par un modèle linéaire en fonction de la moyenne et de la variance d'un ensemble de prévisions.

Pour fournir des prévisions probabilistes, nous proposons une approche innovante fournissant un mélange de distributions [TMB16], au chapitre 3. L'originalité de notre technique provient de l'utilisation de règles de mise à jour des poids issues de l'agrégation séquentielle pour minimiser le CRPS de la distribution empirique pondérée. Grâce à l'utilisation de l'agrégation séquentielle, nos prévisions bénéficient de garanties théoriques de robustesse. L'agrégation séquentielle a déjà été appliquée avec succès aux prévisions quantiles [GGN16 ; BP11], mais pas pour fournir directement une prévision en loi, jusqu'aux travaux de cette thèse [TMB16], et de Baudin [Bau15] et Zamo [Zam16]. Les relations entre les prévisions quantiles et les prévisions en loi sont détaillées dans la Section 1.3.

Par ailleurs, notre étude améliore la connaissance du biais du CRPS pour les distributions empiriques. Le score $S(\mathrm{G}^{\mathcal{E}}, y)$ évalue la qualité de la prévision en loi $\mathrm{G}^{\mathcal{E}}$ de l'observation y. Supposons que $\mathrm{G}^{\mathcal{E}}$ dépende de variables aléatoires, alors le score $S(\mathrm{G}^{\mathcal{E}}, y)$ est également une variable aléatoire. Le biais du score $S(\mathrm{G}^{\mathcal{E}}, y)$ est la différence entre le score moyen $\mathrm{E}(S(\mathrm{G}^{\mathcal{E}}, y))$ et le score de la distribution moyenne $S(\mathrm{E}(\mathrm{G}^{\mathcal{E}}), y)$. Nous généralisons les résultats du biais du CRPS calculé pour un ensemble uniforme de prévisions [FRW08], et nous proposons une minimisation juste du CRPS, sans hypothèse sur des distributions sous-jacentes. Cela est possible en groupant les prévisions par classes de

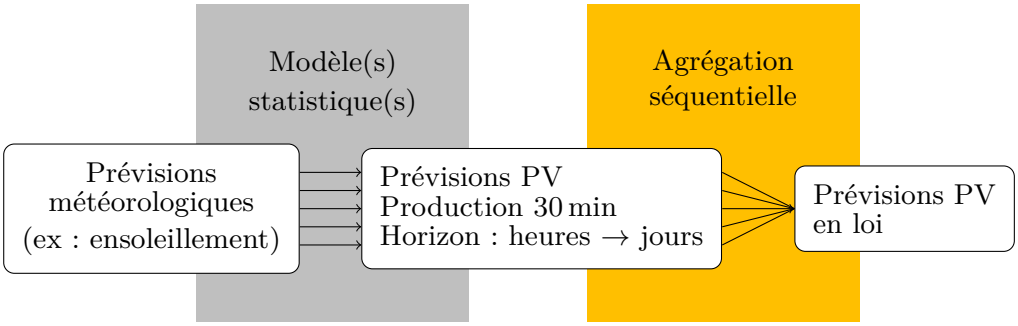


FIGURE A – Description de notre approche des prévisions PV en loi.

membres échangeables, de façon assez similaire au travail de FRALEY et al. [FRG10] pour le BMA.

L'extension des résultats mentionnés ci-dessus est proposée au chapitre 4 pour des pertes, autres que le CRPS, pouvant se décomposer par seuil ou par quantile. De telles décompositions sont étudiées récemment par GNEITING et RANJAN [GR11] et EHM et al. [Ehm+16]. Le sujet n'est toutefois pas nouveau pour l'évaluation des prévisions d'évènement binaire [SAE66]. La question de l'incertitude des observations est abordée à la section 4.2. En assimilation de données, un compromis est généralement trouvé entre de nouvelles observations et une ébauche en fonction de leurs niveaux de confiance respectifs. Des observations possédant un fort niveau de bruit ne seront que peu prises en compte dans la mise à jour du vecteur d'état. Le cas des observations bruitées est également étudié par YANG [Yan04] en agrégation séquentielle. La perte quadratique est normalisée par le niveau de bruit des observations, et une borne de regret par rapport à l'espérance des observations est obtenue sur les pertes simples et sur les pertes normalisées. Notre cadre est différent puisqu'il concerne la prévision probabiliste. Notre objectif premier n'est pas d'obtenir des bornes de regret en espérance, mais de prévoir au mieux la réalité en délivrant une prévision probabiliste. Au lieu de recevoir une observation sans bruit (la réalité), nous considérons que le prévisionniste reçoit de multiples observations ou une distribution. Nous développons cette idée par une technique de moindres carrés généralisés appliquée au CRPS. Des liens sont établis avec des statistiques de test.

Nos techniques sont testées sur des cas d'étude de prévisions en loi pour le PV décrites aux chapitres 5, 6 et 7. Notre approche est résumée par la figure A. Tout d'abord, nous fournissons des prévisions pour 219 parcs de production PV situés en France à la résolution temporelle de 30 min, jusqu'à un horizon de 6 jours. Dans ce cas d'étude, nous utilisons la prévision déterministe HRES et la prévision d'ensemble ENS du ECMWF, Arpège (la prévision déterministe de Météo France) et PEARP (la prévision d'ensemble issue de Arpège). Nous étudions la prévision probabiliste issue des modèles statistiques opérationnels de conversion entre les données de production et les prévisions météorologiques. Nous comparons les performances des ensembles de prévisions et des prévisions quantiles issues de HRES et Arpège. De plus, nous montrons une amélioration de la qualité des prévisions grâce à l'agrégation séquentielle des prévisions de production.

Dans un second cas d'étude, nous étudions l'utilisation du système de prévision AROME (prévision Météo France à haute résolution). La haute résolution spatio-temporelle de AROME permet de générer de multiples prévisions. En plus des incertitudes météorologiques, plusieurs modèles statistiques sont construits pour tenir compte de la difficulté de convertir des prévisions météorologiques en prévisions de production. Grâce à l'agrégation séquentielle, nous montrons qu'il est possible de fournir une prévision de production calibrée issue des prévisions AROME.

Les systèmes de production photovoltaïque insulaires de La Réunion et de la Corse sont étudiés dans un troisième cas d'étude, pour des horizons de prévision courts entre 30 min et 4 h. Nous explorons alors la possibilité de mises à jour infra-journalières grâce à la construction de prévisions prenant en compte toute l'information disponible de la journée. Nous travaillons avec les prévisions météorologiques usuelles ainsi qu'avec des prévisions issues de données satellitaires estimant le mouvement des masses nuageuses. Ces prévisions satellitaires se montrent particulièrement intéressantes pour des horizons de prévision inférieurs à 2 h, alors que les prévisions météorologiques usuelles sont plus utiles pour les horizons de prévision plus longs. Dans tous nos cas d'étude, nous montrons que nos techniques améliorent la qualité des prévisions selon des scores d'évaluation déterministes et probabilistes.

State of the art and contributions

Improved photovoltaic (PV) power integration needs better power forecasts. Forecasters may pursue efforts to improve meteorological models, weather-based power models or statistical post-processing methods. For our part, we focus on the following case: a forecaster, willing to provide probabilistic PV power forecasts, retrieves meteorological forecasts (possibly from various sources). In this general setting, numerous state-of-the-art methods can be tested and combined. Weather forecasts are the basis of our post-processing methods, involving ensembles of forecasts. State-of-the-art statistical models for PV are described in Bacher et al. [BMN09] and Inman et al. [IPC13]. Also, the recent state-of-the-art review [Ant+16] includes a review of probabilistic forecasting and ensemble forecasting for PV applications. Only few publications mention post-processing methods with an ensemble of weather forecasts for PV [Zam+14; Ale+15; SAM16]. Interestingly, Lorenz et al. [LKH12] combines weather forecasts and short-term prediction from satellite data.

A forecaster having multiple forecasts hopefully wishes to combine them in an optimal way. To do so with online learning, also called sequential aggregation, the forecaster gives a weight to each forecast and delivers the combination of the forecasts. The combination rules stemming from online learning depend only on the available past information of each forecast step and come with theoretical performance guarantees under essentially no assumptions (concerning prior weights, underlying stochastic processes or distributions), see the monograph Cesa-Bianchi and Lugosi [CL06], or the gentle reviews Shalev-Shwartz [Sha11] or Stoltz [Sto10] in French. The theoretical performance guarantees are essentially robustness guarantees (regret bounds), ensuring that the combination of the forecasts performs on the long run at least as well as the best forecast or the best fixed combination of forecasts. These techniques have already been tested for several applications: electricity consumption, ozone concentration, wind and geopotential fields [Sto10; MSM09; Mal10; Bau15; GGN16].

We show that online learning may be used to improve solar irradiance forecasts in Chapter 2 based on Thorey et al. [Tho+15]. In this paper, maps of surface solar irradiance are forecasted using ensembles of forecasts from the THORPEX Interactive Grand Global Ensemble (TIGGE) with a 6-h timestep. First, we study ensemble forecasts from 6 meteorological centers: China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), MetOffice (UKMO), Korea Meteorological Administration (KMA), Centro de Previsao Tempo e Estudos Climaticos (CPTEC), and Météo France (M.-F.), providing a total of 158 forecasts. The forecasts are compared with observations derived from MeteoSat Second Generation (MSG) and provided by the HelioClim-3 database as gridded observations over

metropolitan France. Satellite estimations of our data set have a resolution of $1/12^\circ$ which is much finer than the 0.25° resolution at which the forecasts are retrieved. We find that the ensemble forecasts are generally under-dispersed, even when grouped together. Secondly, we use online learning techniques to linearly combine all 158 forecasts with weights that vary in space and time. For each grid point i and for each time t, we apply a discounted ridge regression to find the weights $\boldsymbol{u}_t^{(i)}$ minimizing a certain distance between the past forecasts $\boldsymbol{x}_{m,t'}^{(i)}$ and the past observations $y_{t'}^{(i)}$. The weights $\boldsymbol{u}_t^{(i)}$ are chosen as the minimizer of

$$J(\boldsymbol{u}) = \lambda \|\boldsymbol{u} - \boldsymbol{w}_{\text{ref}}\|_2^2 + \sum_{t'=1}^{t-1} \left(1 + \frac{\gamma}{(t-t')^2}\right) (y_{t'}^{(i)} - \boldsymbol{u} \cdot \boldsymbol{x}_{t'}^{(i)})^2 \,,$$

where $\boldsymbol{w}_{\text{ref}}$ is a reference vector, λ is the regularization parameter and γ is the discount parameter. This method decreases the forecast error by 20% compared to the reference forecast HRES from ECMWF. Besides, this method also corrects the spatial local biases and produces a more realistic spatial pattern of predicted irradiance.

In the meteorological literature, minimizing the CRPS is a common strategy to calibrate probabilistic forecasts [Gne+05; Sch14; JMA15]. However, standard techniques do not offer theoretical guarantees of robustness and usually resort to strong assumptions on the distributions. For example, Bayesian model averaging (BMA) techniques provide a mixture of parametric distributions, usually a Gaussian sum [Gne+05] or gamma distributions sum for wind and precipitation applications [SGR10; Slo+07]. Non-homogeneous regression fits the parameter of a parameterized distribution using characteristics of the ensemble of forecasts [Gne+05; Wil09; TG10]. For instance, a Gaussian distribution is fitted using a linear model between the mean of the distribution and the mean of the forecasts. Likewise the standard deviation of the Gaussian distribution is fitted according to the ensemble spread.

To provide probabilistic forecasts, we propose an innovative approach by combining multiple forecasts in a linear opinion pool [TMB16], in Chapter 3. The originality of our technique is to use combination rules deriving from online learning techniques in order to minimize the CRPS of the weighted empirical distribution function. Because we use online learning techniques, our forecasts come with theoretical guarantees of robustness. Online learning techniques have already been applied successfully to quantile prediction [GGN16; BP11], but not for directly delivering probabilistic forecasts, until this thesis [TMB16], and the works of Baudin [Bau15] and Zamo [Zam16]. Relationships between quantile losses and probabilistic forecasts are detailed in Section 1.3.

Our study also improves the knowledge of the CRPS expectation for linear opinion pools. The score $S(\mathrm{G}^{\mathcal{E}}, y)$ evaluates the quality of the probabilistic forecast $\mathrm{G}^{\mathcal{E}}$ for the observation y. Assume $\mathrm{G}^{\mathcal{E}}$ relies on random variables, the score $S(\mathrm{G}^{\mathcal{E}}, y)$ is also a random variable. The bias of the score $S(\mathrm{G}^{\mathcal{E}}, y)$ is the difference between the average score $\mathrm{E}(S(\mathrm{G}^{\mathcal{E}}, y))$ and the score for the average distribution $S(\mathrm{E}(\mathrm{G}^{\mathcal{E}}), y)$. We generalize results on the bias of the CRPS computed with ensemble forecasts [FRW08], and propose a new scheme to achieve fair CRPS minimization, without any assumption on the distributions. This is achieved by grouping forecasts in classes of exchangeable

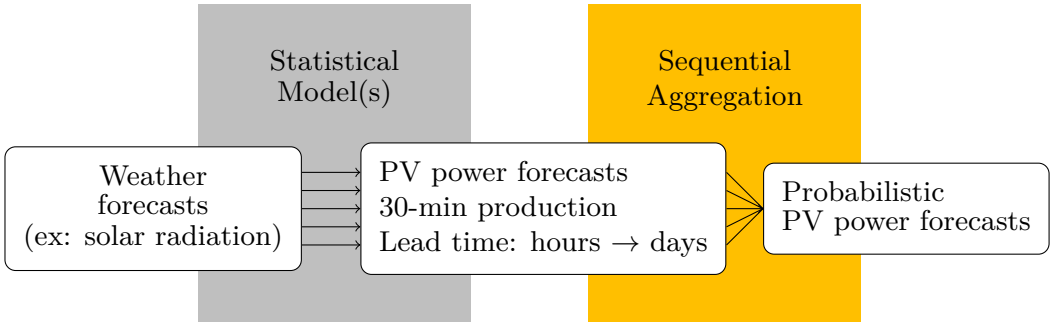


Figure B – Description of our approach to PV power forecasts.

members, quite similarly to the work of Fraley et al. [FRG10] for BMA.

Extension of the above results is proposed in Chapter 4 to other losses than the CRPS, which admit a threshold decomposition or a quantile decomposition. Such decomposition was recently studied by Gneiting and Ranjan [GR11] and Ehm et al. [Ehm+16]. This topic is not new for the evaluation of binary event forecasts [SAE66]. The question of uncertainty in the observations is addressed in Section 4.2. In data assimilation, a trade-off is commonly achieved between observational information and the background state based on their respective level of confidence. Observations with a high level of noise will not have much importance for the state update. Observational noise is studied by Yang [Yan04] for online learning. They normalize the square loss by the observation noise rate and obtain regret bounds on the normalized and unnormalized losses computed against the distribution mean. Our work is slightly different since we intend to provide probabilistic forecasts. Our first objective is not to obtain regret bounds in expectation, but to forecast as best as possible the reality by delivering a probabilistic forecast. Instead of receiving one single noiseless observation (the reality), we consider that the forecaster receives multiple observations or a distribution. We develop this idea using a generalized least-square method applied to the CRPS and provide connections to test-statistics.

Our new techniques are tested in several case studies of probabilistic PV power forecasts described in Chapters 5, 6 and 7. Our approach is summarized in Figure B. First, we provide forecasts for 219 photovoltaic (PV) power plants located in France with a 30-min timestep, up to 6 days of lead time. In this case study, we use ECMWF deterministic forecast HRES and ensemble ENS, Arpège (Météo France deterministic forecast) and PEARP (Météo France ensemble forecast). We investigate PV probabilistic forecasting with the operational statistical models between the production data and the weather forecasts. We compare the predictive performance of ensemble forecasts and quantile forecasts deriving from the deterministic HRES and Arpège. We also provide improvements on the calibration of the probabilistic forecasts by combining the production forecasts.

In a second case study, we investigate the use of AROME forecasts (Météo France high-resolution forecasts). The high spatio-temporal resolution of AROME enables

to generate multiple forecasts accounting for various local scenarios. In addition to weather uncertainties, several statistical models are built to account for the difficulty of converting weather forecasts to production forecasts. Using online learning techniques, we show that we can provide calibrated production forecasts using the single high-resolution forecasting system AROME.

The insular PV power production of Réunion and Corsica are investigated in a third case study focusing on short lead times from 30 min to 4 h. Here we explore the possibility of intraday updates with the generation of new members incorporating all the available information of the day. Namely, new forecasts are built using regular PV forecasts and available production data of the current day as inputs. We work with usual weather forecasts as well as predictions from satellite data derived from cloud-motion analysis. Prediction using satellite data are most useful for short lead times below 2 h, while regular weather forecasts deliver the most useful information for longer lead times. For all the case studies, we show that our technique provides forecast improvements for multiple evaluation tools both deterministic and probabilistic.

1 Introduction

This chapter aims at providing background knowledge on this thesis. First, we introduce photovoltaic power forecasts using weather forecasts, from the industrial context to statistical learning challenges. The theoretical background of sequential aggregation, or online learning with multiple experts, is then detailed. We also delineate our position on probabilistic forecasting and introduce probabilistic forecasts evaluation.

Contents

1.1 Forecasting photovoltaic power production with meteorological forecasts

1.1.1 Context

This thesis aims at improving forecasts of photovoltaic (PV) power production. To be more specific, the company "Electricité de France" (EDF) is particularly interested in forecasting PV power for several geographical areas, in metropolitan France or for insular electric systems. For example, we model an estimation of the total production of Metropolitan France, the total production of Corsica Island and the total production of Réunion.

	Growth 2015	Growth 2016	Installed capacity
Solar Power	+0.9	+0.6	6.8
Wind Power	+1.0	+1.3	11.7
Nuclear Power			63.1
Hydropower			25.5

Table 1.1 – Growth and installed capacity (in GW) for solar, wind, nuclear and hydro power.

The installed capacity of renewable electricity is growing in France. The need for forecasts of renewable energy production grows with the amount of renewable energy delivered to the electrical grid. Over the past decade, the installed capacity of wind and solar power drastically increased. The current annual growth rate is close to 10%, see Table 1.1. Indeed, the installed capacity of solar power increased in France by 899 MW in year 2015 and 576 MW in 2016, to reach the amount of 6.8 GW. Over the same periods, the installed capacity of wind power increased in France by 1011 MW in year 2015 and 1345 MW in 2016, to reach the amount of 11.67 GW. This strong growth is significant compared to the total installed capacity of 129 GW at the beginning of 2016, with above 63 GW of nuclear capacity and above 25 GW of hydropower capacity. These numbers need to be put into perspective, because wind and solar are intermittent energy sources. One gigawatt of PV power is not equivalent to one gigawatt of nuclear power, at least in terms of load factor. The load factor is the ratio of the produced energy and the energy that would have been produced if the production level was maximal. In fact, the load factor of wind and solar power are quite low compared to the load factor of nuclear power plants (around 75%). The average load factor of solar power is of 15% with a strong seasonality (6% in December and 20% in summer). Besides, the load factor of wind power is around 24% (30-33% in winter and 15-20% in summer). We refer the interested reader to the periodic report "Overview of Renewable Electricity" [*] (in French "Panorama de l'électricité renouvelable"), and to the "Forecast assessment of electricity supply-demand balance" [†] (in French "Bilan prévisionnel de l'équilibre offre-demande d'électricité en France"), both published by the French Transmission System Operator "Réseau de Transport d'électricité" (RTE).

What is at stake in renewable energy forecasting ?

The production forecasts are achieved at various geographical and temporal scales, depending on the production target. For example, the production can be local at the scale of a production unit, or at a larger geographical scale, i.e. a sum of local productions.

For operational purposes, Transmission System Operators and Distribution System Operators need production forecasts to anticipate constraints on the electrical grid, generated by supply-demand imbalances. For financial purposes, Balance Responsible Entities (BR) also need production forecasts. A BR must declare its injections and extractions of power on the grid, with constraints on its so-called balance perimeter.

∗. http://www.rte-france.com/sites/default/files/panorama_enr20161231.pdf
†. http://www.rte-france.com/sites/default/files/bp2016_complet_vf.pdf

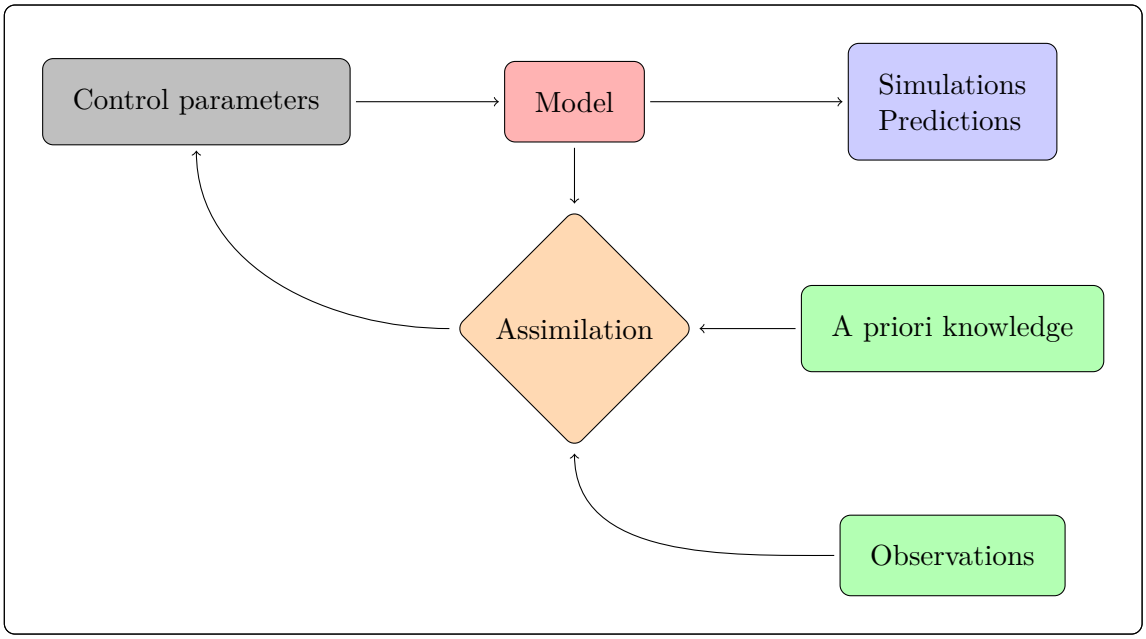


Figure 1.1 – Illustration of data assimilation.

These constraints ensure the whole system to be balanced. When injections and extractions do not match the BR declaration, the BR suffers financial penalties. The BRs can trade electricity on the market to ensure the balance of its perimeter or take benefits from the market price. Therefore, improved production forecasts on their perimeter means better market positioning and avoided penalties for BRs. Historically in France, EDF was the main BR with solar and wind production inside its balance perimeter due to the mechanism of feed-in tariffs. The trends of the market evolution is that wind and solar energy producers will introduce their own product on the market, or sell it directly to a BR.

Renewable energy forecasts can also be integrated to maintenance planning. The motivation behind is that the maintenance should be done

— when the production is low for renewable energy producers,

— without generating constraints on the grid for grid operators.

1.1.2 Weather forecasting

Wind and solar power productions are strongly dependent on the weather, hence wind and solar forecasts resort to numerical weather predictions (NWPs) for lead times between a few hours to several days.

How numerical weather predictions are generated ? NWPs are produced with data assimilation schemes, the most famous being the Kalman filter [Kal60]. Data assimilation integrates several elements: a state vector, a model, a priori knowledge, observations, an observation function and control parameters, as represented in Figure 1.1. These elements are not perfect and possibly corrupted by noise. The purpose

	AROME	Arpège	PEARP	HRES	ENS
Spatial resolution	1.3 km	7.5 km	15.5 km	0.1°	0.2°
Temporal resolution	1 h	1 h	3 h	3 h	3 h
Lead time	2 days	4 days	4 days	10 days	15 days

Table 1.2 – Meteorological forecasts and their resolutions over Europe in 2016. The lead time is indicated for the forecast of the 00 UTC analysis. The temporal resolution of PEARP is only of 6-h after the lead time of 48 h, and the temporal resolution of ECMWF forecasts is of 6-h after the lead time of 150 h. The conversion from degrees to kilometers is roughly done with a factor 100.

of data assimilation is the best evaluation of the state vector with the available information. After the reception of new observations, the state vector is updated to take into account this new information, given characteristics on observations errors, background errors on the state and model errors.

For numerical weather predictions:

— The state vector describes the atmospheric state and includes several fields such as horizontal wind speed and temperature.

— Atmospheric models are built from partial differential equations representing the atmospheric motions and many simplifications (or parameterizations), which tend to model the effect of a process rather than the process itself. The model is run from a given state to create predictions of the evolution of this state.

— The observation function relates the state space to the observation space. Indeed, the observations generally do not live in the same space as the state. For example, satellite images are used as observations, but do not always have direct counterparts in the model state.

— The size of the observation and state vectors are huge. They easily reaches 10^6 for the observations and 10^9 for the state vector, hence computational problems arise.

Numerical weather predictions used in this thesis.

In practice, several meteorological forecasting systems are used in this thesis. Their current characteristics are summarized in Table 1.2. Besides the model and the data assimilation scheme, the forecasts may differ in their covered geographical areas, their lead time, and their spatial and temporal resolutions. Ensemble forecasts are traditionally generated with the same model as a deterministic forecast, but at a coarser resolution with slight changes in the model physics, in the parameterizations, or with different perturbations of the initial conditions. Therefore ensemble forecasts integrate different sources of errors and quantify the forecast uncertainty. All along this thesis, we show the benefits of using several forecasts. This approach begins with the variety of weather forecasts that we use as inputs.

Météo France generates the global model Arpège, from which the 34 ensemble forecasts PEARP are derived. Météo France also provides mainly for France the local model AROME, designed to forecast severe weather events such as heavy rains in the south of France. The resolutions of AROME are quite fine (currently 1.3 km, 1 h), at

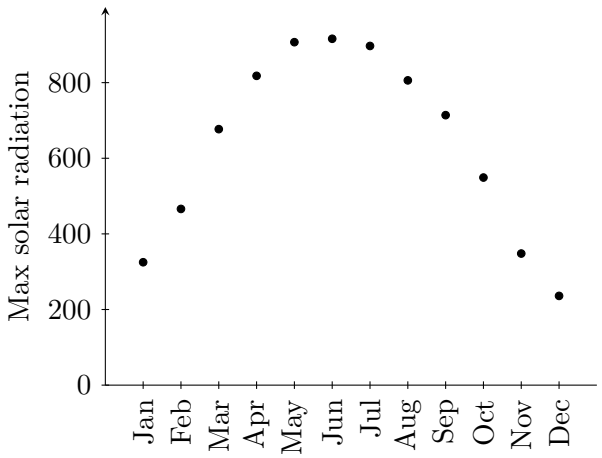


Figure 1.2 – Monthly maximum solar radiation (in $\mathrm{W\,m^{-2}}$) in 2012 for one location, according to HRES. We recall that HRES data are 3-h averages.

the cost of limited geographical extension and limited lead time of two days. We also use atmospheric forecasts from ECMWF: the deterministic forecast HRES and the ensemble of forecasts ENS. For solar forecasting with a 6-h temporal resolution, TIGGE (THORPEX Interactive Grand Global Ensemble) ensembles from several meteorological centers were used. The detailed description of TIGGE ensembles is postponed to Chapter 2.

1.1.3 Solar radiation forecasting

For PV applications, the main meteorological fields are irradiance (or solar radiation) and 2-m temperature. Irradiance is directly related to the energy produced by the solar panels and high temperature at ground-level decreases the panel efficiency.

The position of the Sun compared to the position of the Earth controls directly the top-of-atmosphere irradiance. Along one year, the evolution of these positions generates the diurnal and the seasonal cycles. The diurnal cycle facilitates solar forecasting, since we know that every day begins with a null level of radiation. For a clear sky day, the main challenge is to estimate the bell-shape of solar radiation and the maximal reached value. The seasonal cycle is clearly depicted with monthly maximal value of solar radiation, see Figure 1.2. The ratio between the maximal value of January and June almost reach a factor 4. We highlight the fact that solar radiation data are time-averaged, and usually indicate the average flux over the past period. For example, solar radiation at 12:00 with a 3-h temporal resolution is the average flux between 9:00 and 12:00.

From top-of-atmosphere to ground-level, solar radiation encounters many processes (absorption, scattering, reflection) mainly occurring in clouds. Hence irradiance maps are closely related to nebulosity maps. Cloud cover and nebulosity are often used as a proxy for solar radiation. Solar radiation maps of HelioClim (real-time estimations from satellite data), AROME and HRES forecasts (12-h lead time) are shown in Figure 1.3, for 2013-06-13 at 12:00 (UTC). Note that HelioClim and AROME maps are 1-h averages

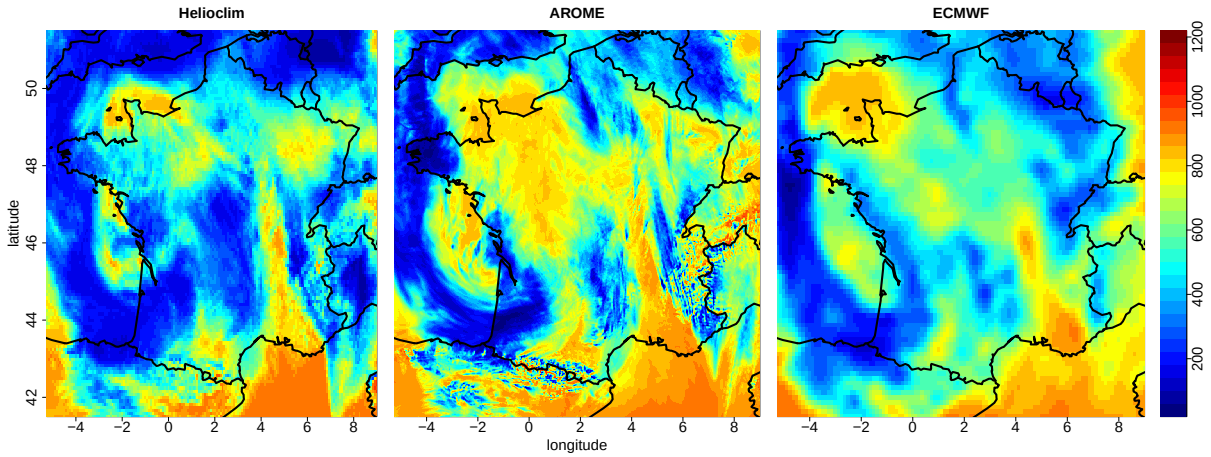


Figure 1.3 – Maps of solar radiation for 2013-06-13 at 12:00 (UTC): HelioClim real-time estimation (left), AROME forecast (center), HRES forecast (right).

while the HRES map is a 3-h average. The spatial variations of irradiance are quite high since very cloudy areas may receive less than one third of the energy received by clear-sky areas. These variations are seen by the forecasts at the scale of a few tens of kilometers, or even less. The tremendous impacts of orographic effects in mountains areas with many scattered clouds can be seen in Figure 1.3.

We see that spatial resolution is a key-issue in solar forecasting. At the scale of a grid-point, one may ask whether the level of irradiance is correctly forecasted. However, at the scale of a regional map, one may ask whether the clouds are modeled at the correct time and location. Indeed, a cloud may be "seen" by the model, but located with an error of 50 km. As a conclusion, we note that:

— Observations may be local (ground-station measurements) or estimated with satellite data. Satellite data provide a better spatial representation, but may be more prone to calibration errors than local measurements.

— We need a validation criterion to assess the quality of a forecast. This validation criterion may depend on the spatial scale of the validation.

— The forecasted average over several grid points may be the best estimation for comparisons at the grid-point scale.

Using high-resolution forecasts such as AROME is a delicate task, see our contribution in Chapter 6.

1.1.4 PV power data and statistical modeling

As opposed to irradiance maps, production data of a PV power plant are localized. For lead times of several hours to several days, we usually work at a resolution of 30 min or 1 h. Averaging in time production data smooths the variability of the data, as shown in Figure 1.4, and facilitates the work of the forecaster. Indeed, the local production peaks are barely seen after time-averaging. The forecaster faces a dilemma when the production data show a large uncertainty. Should they attempt to forecast the average,

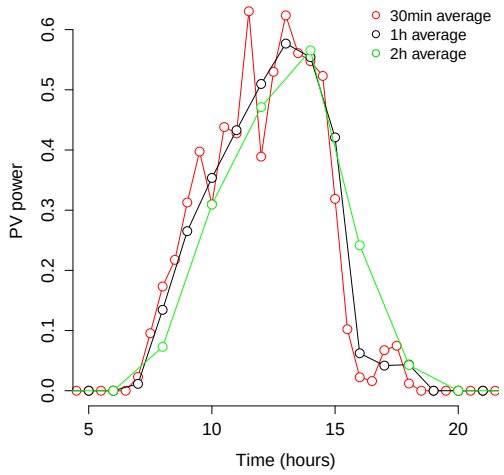


Figure 1.4 – Production data of one plant, normalized by the installed capacity. The temporal averages for 30-min, 1-h and 2-h resolutions are shown.

or should they attempt to forecast local peaks although the prediction may not be at the correct time ?

Weather forecasts are converted to PV forecasts with statistical or physical models. Besides weather variables, PV power production rely on the technology of the solar panels and on the incident angle of solar radiation on the solar panels. Hence physical models include the panels orientation and inclination. Solar radiation data are a good starting point to build a statistical model. Using solar radiation estimates from HelioClim (and not forecasts), we see in Figure 1.5 that linear models are a good approximation of the relationship between solar radiation and PV production, for a period of nearly 20 days. Here we simply applied a multiplicative factor to the solar radiation estimates. Advanced models often use clear-sky production profiles and clear-sky radiation profiles computed for each day of the year. In this setting, weather forecasts provide indications on possible production decreases along the day. The production P and solar forecasts I are converted respectively to the clear-sky indices τ_P and τ_I:

$$P = \tau_P P_{cc} \quad \text{and} \quad I = \tau_I I_{cc}\,,$$

where clear sky production P_{cc} and clear sky solar radiation I_{cc} are the production and solar radiation in clear sky conditions. A statistical regression estimates the parameters a_i between weather forecasts and the forecasts $\widehat{\tau_P}$ of clear-sky index:

$$\widehat{\tau_P} = a_0 + a_1\tau_I + a_2\tau_I^2 + a_3 Tcc\,.$$

In this example, the non-linear term τ_I^2 and the total cloud cover Tcc are added as explanatory variables. The statistical models of our PV forecasts are detailed in Chapter 5.

In practice, the forecaster encounters major difficulties:

— The forecaster cannot determine the best set of parameters for the current period, but only for a past period. This may result in drifting effects or inaccurate parameter estimates due to inter-year variability.

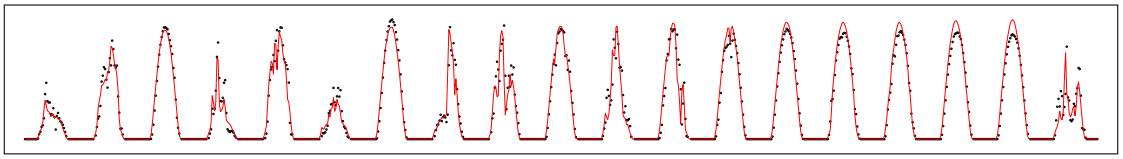


Figure 1.5 – Time-series of Helioclim (red line) and production data (black dots), with simple scaling of the data (for one plant in April 2013).

— Solar radiation forecasts may be quite far from the reality. Systematic biases such as seasonal biases may be corrected with statistical models taking into account the hour of the day, the time in the year and the lead time of the forecasts. However, specializing the statistical model increases the amount of model parameters, or reduces the available amount of data to fit several models.

— PV power observations are not always available, due to data corruption or lack of metering.

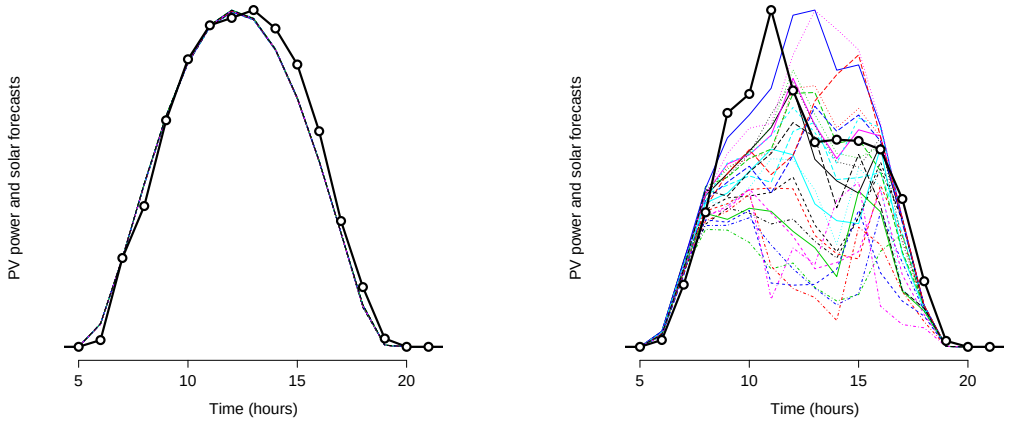
Interestingly, the forecaster can use statistical methods as downscaling methods. For example, statistical models are built between forecasts at a 3-h resolution and observations at a 30-min resolution in order to provide forecasts at the 30-min resolution. Thanks to the fine resolution of the observations, it is possible to go under the resolution of the forecasts. The same principle is applied by using HelioClim as solar radiation observations to improve the spatial resolution of the forecasts in Chapter 2.

1.1.5 Probabilistic forecasts of PV power with meteorological forecasts

Using multiple grid points gives a good insight into the spatial variability in the irradiance maps, see Figure 1.6. We show production data and solar forecasts from AROME, normalized by the daily maximum value. In a clear-sky situation (1.6(a)), the forecasts of nearby grid-points are almost identical. But in a situation with large uncertainties on the cloud cover (1.6(b)), the spatial variability of the solar forecasts is very high. For this particular day, the information of cloud cover variability is correct and noticeable in the production data. Our point here is to illustrate that using only one single forecast may not be appropriate, and that solar forecasts integrate complex spatio-temporal information.

Probabilistic forecasts acknowledge the inability of the forecaster to provide a priori a perfect estimate of the observation, that we consider here noiseless. Probabilistic forecasts are conveniently described by a Cumulative Distribution Function (CDF). They give the probability (according to the forecaster) that the observation is below a threshold. Confidence intervals are derived from the CDF of the forecaster. For example, the forecaster says that there is 5% chance that the production is below the number a and 95% chance that the observation is below the number b. The interval $[a, b]$ is then a 90% confidence interval.

We illustrate probabilistic forecasts for PV applications with a 30-min temporal resolution in Figure 1.7. For two consecutive days, the forecaster delivers a probabilistic forecast and a deterministic forecast. The median of the probabilistic forecasts mainly



(a) 2013-04-15 (b) 2013-04-08

Figure 1.6 – Production data (black line with white dots) and local solar radiation forecasts (5×5 nearby grid points from AROME). The data are normalized by the daily maximal value. For the clear day situation (left), the solar forecasts of the 5×5 nearby grid points are equal.

indicates a change in the level of production between the two days. With the median only, one cannot know whether a large intraday variability is expected. This information is clearly given by the confidence intervals, which are much larger for the first day than for the second day. In this example, the observed production has indeed a larger intraday variability during the first day. The large confidence intervals reflect the inability of the forecaster to describe precisely the observation variations. Besides, the deterministic forecast provides additional information, such as production decreases during the first day. The integration of this information is possible thanks to sequential aggregation, as described in Section 1.2.

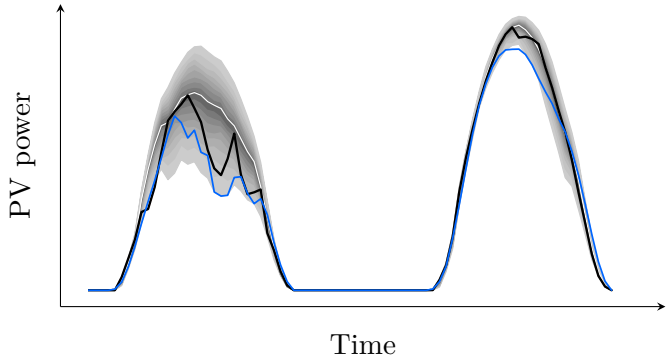


Figure 1.7 – PV production (black), deterministic forecast (blue) and probabilistic forecasts (confidence intervals in shaded gray, median in white) for two consecutive days.

1.2 Sequential aggregation

1.2.1 Context

In this section, we introduce online learning in the expert setting. Just before time $t > 0$, a forecaster aims to deliver the best possible prediction for the observation $y_t \in \mathcal{Y}$, while an advice of M experts $x_{m,t} \in \mathcal{X}$ is given to the forecaster. In other words, the forecaster receives several forecasts (or experts) and wishes to make an optimal use of this information. In real-world applications, the forecaster may also build experts and not only receive them. The key-point of sequential aggregation is nevertheless to ensure performance guarantees, given a set of arbitrary forecasts.

We work in the deterministic setting of individual sequences : no assumptions are made on the observations y_t or on the experts $x_{m,t}$. This setting is particularly interesting, because the performance guarantees described below are ensured no matter the received data (observations and forecasts). The methods providing performance guarantee with individual sequences are therefore intrinsically robust, see Cesa-Bianchi and Lugosi [CL06] for an in-depth analysis.

In practice, the forecaster gives the weight $u_{m,t}$ to the expert $x_{m,t}$, and provides the forecast $s_t(\boldsymbol{u}_t, \boldsymbol{x}_t)$, where the function s_t can be arbitrarily chosen by the forecaster. A simple combination of forecast $s_t(\boldsymbol{u}_t, \boldsymbol{x}_t) = \boldsymbol{u}_t^\top \boldsymbol{x}_t = \sum_{m=1}^M u_{m,t} x_{m,t}$ is often chosen. The general algorithm reads:

Initialization: $\boldsymbol{u}_1$;

For each time index t = 1, 2, ..., T

1. get the vector of predictions data $\boldsymbol{x}_t$,
2. compute the forecaster's choice $s_t(\boldsymbol{u}_t, \boldsymbol{x}_t)$,
3. get the verification y_t and compute $\boldsymbol{u}_{t+1}$, based on the update rule.

The initial weight vector $\boldsymbol{u}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$. Several examples of update rules are given in Section 1.2.3. We highlight here the fact that $\mathcal{X}$ and $\mathcal{Y}$ may be functional spaces. Hence it is possible to combine probabilistic forecasts (or CDFs) as described in Section 1.2.4.

1.2.2 Algorithm evaluation with regret bounds

The forecaster is willing to make the best possible forecast, where "best" refers to a real-valued loss function ℓ, measuring the distance between the observation and the forecast. The loss ℓ is negatively oriented, meaning that it should be as low as possible. The loss $\ell_t(\boldsymbol{u})$ for time t is defined with implicit relationship with the experts $x_{m,t}$ and the observation y_t. This notation is used because the weight vector $\boldsymbol{u}$ is the forecaster's choice while the experts and observations are arbitrarily given to the forecaster.

The expression "best possible forecast" introduces the notion of algorithm evaluation. The total loss of the forecaster may be divided into two terms:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) = \underbrace{\inf_{\boldsymbol{u}\in\mathcal{U}} \left[\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \right]}_{\text{approximation error}} + \underbrace{\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{U}} \left[\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \right]}_{\text{estimation error}},$$

of approximation and estimation errors. The approximation error is the error of the best possible forecast with fixed weights in $\mathcal{U}$. While the approximation error is considered as being unavoidable, the estimation error should be as low as possible. The estimation error is referred to as the cumulated regret of the forecaster, who competes against the best combination of experts. The cumulated regret of an algorithm $\mathcal{A}$, generating the weights $\boldsymbol{u}_t$, is defined by

$$R_T(\mathcal{A}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{U}} \left[\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \right] .$$

Standard algorithms show a sublinearity of R_T in T, expressed by $\lim_{T\to\infty} R_T \leqslant o(T)$. This guarantees that the combination of the experts defined by the weights $\boldsymbol{u}_t$ is asymptotically at least as good as the best fixed combination of experts. Consequently, the combination of experts is also asymptotically at least as good as the best experts and the best fixed subset of experts with uniform weights.

1.2.3 Online learning algorithms

In this section, we introduce several algorithms to find the weights $\boldsymbol{u}_t$. The update rules are often described by a regularized regression. The weight vector $\boldsymbol{u}_t$ is then the minimizer of a cost function, with general formulation

$$\boldsymbol{u}_t = \arg\min_{\boldsymbol{u}\in\mathcal{U}} \left[\underbrace{\Psi_{0,t}(\boldsymbol{u}) + \sum_{s=1}^{t-1} \Psi_{s,t}(\boldsymbol{u}, \boldsymbol{u}_s)}_{\text{Regularization}} + \underbrace{\sum_{s=1}^{t-1} \widetilde{\ell}_s(\boldsymbol{u})}_{\text{Loss}} \right]$$

To obtain a general expression, we noted above $\widetilde{\ell}_s(\boldsymbol{u})$ instead of $\ell_s(\boldsymbol{u})$ to indicate that the terms $\ell_s(\boldsymbol{u})$ are eventually replaced by a linear or a quadratic approximation $\widetilde{\ell}_s(\boldsymbol{u})$ of $\ell_s(\boldsymbol{u})$, for example around u_s [KW97; HAK07; McM11]. The losses $\widetilde{\ell}_s(\boldsymbol{u})$ are to be minimized with a trade-off against the regularization terms $\Psi_{s,t}$ to control the variations of $\boldsymbol{u}_t$ from the previous weights $\boldsymbol{u}_s$ or from a reference weight vector $\boldsymbol{u}^{\text{ref}}$ (or prior) implicitly defined in $\Psi_{0,t}$. For example, regularization terms can be of 2-norm type $\|\boldsymbol{u} - \boldsymbol{u}^{\text{ref}}\|_2^2$, of Kullback-Leibler type $\sum_m u_m \ln(\frac{u_m}{u_m^{\text{ref}}})$ or of χ^2 type $\sum_m \frac{(u_m - u_m^{\text{ref}})^2}{u_m^{\text{ref}}}$ [KW97]. The χ^2 regularization may be seen as a second-order approximation around $\boldsymbol{u}^{\text{ref}}$ of the Kullback-Leibler regularization.

Ridge Regression. A celebrated example is the ridge regression for the square loss $((\boldsymbol{u}^\top \boldsymbol{x}_t) - y_t)^2$. The weight vector $\boldsymbol{u}_t \in \mathbb{R}^M$ is chosen as the minimizer of

$$J(\boldsymbol{u}) = \lambda \|\boldsymbol{u}\|_2^2 + \sum_{t'=1}^{t-1} (y_{t'} - \boldsymbol{u}^\top \boldsymbol{x}_{t'})^2 .$$

The parameter λ controls the trade-off between the losses and the 2-norm regularization. The ridge regression has the following regret bound:

$$\sum_{t=1}^{T} (y_t - \boldsymbol{u}_t^\top \boldsymbol{x}_t)^2 - \min_{\boldsymbol{u}\in\mathcal{B}_M} \sum_{t=1}^{T} \left(y_t - \boldsymbol{u}^\top \boldsymbol{x}_t \right)^2 \leqslant \mathcal{O}\left(\ln T\right) ,$$

under the assumption of bounded losses $(y_t - \widehat{y}_t)^2$. The weight $\boldsymbol{u}$ is searched in a 2-norm ball $\mathcal{B}_M$ in $\mathbb{R}^M$. The assumption of bounded losses may not always hold, but this difficulty is overcome with an additional regularization term $(\boldsymbol{u}^\top \boldsymbol{x}_t)^2$ [Vov01; AW01]. We refer the reader to the introduction of Gerchinovitz [Ger11] for a concise analysis. Still, key points concerning the ridge regression are the logarithm regret bound, and its quadratic formulation. From an optimization perspective, the ridge regression is a very handy and powerful tool. A modified version of the ridge regression is used in Chapter 2.

Exponentiated Weighted Average. A most standard algorithm, and among the oldest, is the Exponentiated Weighted Average (EWA) [LW94]. This algorithm uses linear losses $\sum_{m=1}^M u_{m,t}\ell_{m,t}$ with positive weights summing to one (defining the simplex $\mathcal{P}_M$), and bounded losses $\ell_{m,t} \in [a,b]$. In EWA, the regularization is the Kullback-Leibler divergence of $\boldsymbol{u}_t$ from $\boldsymbol{u}$. The update rule of EWA with learning rate η reads

$$\boldsymbol{u}_{t+1} = \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{P}_M} \sum_{m=1}^M u_m \ln\left(\frac{u_m}{u_{m,t}}\right) + \eta u_m \ell_{m,t}\,,$$

giving the update rule

$$u_{m,t+1} = \frac{u_{m,t}\exp\left(-\eta\ell_{m,t}\right)}{\sum_{m'=1}^M u_{m',t}\exp\left(-\eta\ell_{m',t}\right)}\,.$$

The weight $u_{m,t+1}$ is derived from $u_{m,t}$ with a multiplicative factor, accounting for the loss $\ell_{m,t}$. The idea is that the weight of a good expert (with a small loss compared to the other losses) increases with time, and the learning rate controls the speed of the variation of $\boldsymbol{u}_t$. Because the losses are linear, the best fixed combination of expert is the best expert: $\inf_{\boldsymbol{u}\in\mathcal{U}}\left[\sum_{t=1}^T\sum_{m=1}^M u_m\ell_{m,t}\right] = \min_k \sum_{t=1}^T \ell_{k,t}$. In other words, the forecaster competes against the best expert. The algorithm EWA admits the following regret bound:

$$\sup\left[\sum_{t=1}^T\sum_{m=1}^M u_{m,t}\ell_{m,t} - \min_k \sum_{t=1}^T \ell_{k,t}\right] \leqslant \frac{\ln M}{\eta} + \eta\frac{(b-a)^2}{8}T\,.$$

The algorithm EWA may also be used with any convex function $\ell(\boldsymbol{u})$, thanks to Jensen's inequality [‡]. Using the bound for linear losses and Jensen's inequality, we have

$$\begin{aligned}\sum_{t=1}^T \ell(\boldsymbol{u}_t) - \min_k\left\{\sum_{t=1}^T \ell_{k,t}\right\} &\leqslant \sum_{t=1}^T\sum_{m=1}^M u_{m,t}\ell_{m,t} - \min_k\left\{\sum_{t=1}^T \ell_{k,t}\right\}\\ &\leqslant \frac{\ln M}{\eta} + \eta\frac{(b-a)^2}{8}T\,.\end{aligned}$$

We see that a limitation of EWA is that the forecaster competes against the best expert, but not again the best fixed combination of experts. The bound is minimized for the

‡. For a convex function ψ and a distribution $\mathcal{D}$, we have $\psi(\mathrm{E}_{z\sim\mathcal{D}}(z)) \leqslant \mathrm{E}_{z\sim\mathcal{D}}(\psi(z))$. Roughly speaking, the value of the expectation is inferior to the expectation of the values.

optimal learning rate $\eta^\star = (b-a)^{-1}\sqrt{8(\ln M)/T}$, and the minimal value of the bounds, reached with $\eta^\star$, is equal to $(b-a)\sqrt{(\ln M)T/2} = o(T)$. A forecaster may use several learning rates according to the length of the experiment in order to circumvent the dependency in T of $\eta^\star$. This is referred to as the doubling trick or its variants [CL06; Bau15].

Exponentiated gradient (EG). Using EWA and the so-called gradient trick, the forecaster can also compete against the best convex combination. The resulting algorithm is Exponentiated Gradient, defined by the update rule

$$u_{m,t+1} = \frac{u_{m,t}\exp\left(-\eta\widetilde{\ell}_{m,t}\right)}{\sum_{m'=1}^{M} u_{m',t}\exp\left(-\eta\widetilde{\ell}_{m',t}\right)}\,.$$

Strictly speaking, EG is built by replacing $\ell_{m,t}$ by the loss gradient

$$\widetilde{\ell}_{m,t} = \frac{\partial \ell_t}{\partial u_m}(\boldsymbol{u}_t)\,,$$

in the update rule of EWA.

The gradient trick uses the convexity of ℓ to obtain:

$$\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \leqslant (\boldsymbol{u}_t - \boldsymbol{u})^\top \nabla\boldsymbol{\ell}_t(\boldsymbol{u}_t) = \boldsymbol{u}_t^\top\widetilde{\boldsymbol{\ell}}_t - \boldsymbol{u}^\top\widetilde{\boldsymbol{\ell}}_t\,.$$

for any two vectors $\boldsymbol{u}_t, \boldsymbol{u} \in \mathcal{P}_M$. Summing over time, we get the following regret bound inequalities:

$$\begin{aligned}
\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}}\sum_{t=1}^{T}\ell_t(\boldsymbol{u}) &= \sup_{\boldsymbol{u}\in\mathcal{P}}\left(\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u})\right)\\
&\leqslant \sup_{\boldsymbol{u}\in\mathcal{P}}\left(\sum_{t=1}^{T}\sum_{m=1}^{M} u_{m,t}\widetilde{\ell}_{m,t} - u_m\widetilde{\ell}_{m,t}\right)\\
&= \sum_{t=1}^{T}\sum_{m=1}^{M} u_{m,t}\widetilde{\ell}_{m,t} - \min_k\sum_{t=1}^{T}\widetilde{\ell}_{k,t}\\
&\leqslant \frac{\ln M}{\eta} + \eta\frac{(\widetilde{b}-\widetilde{a})^2}{8}T\,.
\end{aligned}$$

The regret bound of EWA with the loss gradients gives the last inequality, with the assumptions of bounded gradients $\widetilde{\ell}_{m,t} \in [\widetilde{a}, \widetilde{b}]$.

Polynomially weighted averages with multiple learning rates (ML-Poly). The algorithm ML-Poly [GSE14], described in Table 1.3, enjoys two desirable properties: it has no parameters and adapts to the difficulty of the data. The algorithm relies on the excess loss $\boldsymbol{u}_t^\top\widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$ reflecting the regret of the forecaster to play the weighted combination $\boldsymbol{u}_t$ instead of the mth expert. Designed for linear losses, the algorithm ML-Poly may be used with convex loss functions thanks to the gradient trick. The regret bound of ML-Poly is expressed against the best member for the linearized losses. For all sequences of losses $\widetilde{\ell}_{m,t} \in [0,1]$, the cumulated loss of ML-Poly is bounded:

$$\sum_{t=1}^{T}\boldsymbol{u}_t^\top\widetilde{\boldsymbol{\ell}}_t \leqslant \min_{1\leqslant m\leqslant M}\left\{\sum_{t=1}^{T}\widetilde{\ell}_{m,t} + \sqrt{M(1+\ln(1+T))\left(1+\sum_{t=1}^{T}(\boldsymbol{u}_t^\top\widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2\right)}\right\}\,.$$

update the learning rate of each member	$\eta_{m,t} = 1/\left(1 + \sum_{t'=1}^{t} (\boldsymbol{u}_{t'}^\top \widetilde{\boldsymbol{\ell}}_{t'} - \widetilde{\ell}_{m,t'})^2\right)$
update the regret of each member	$R_{m,t} = R_{m,t-1} + \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$
compute the weights	$u_{m,t+1} = \eta_{m,t}(R_{m,t})_+ \;/\; \boldsymbol{\eta_t}^\top (\boldsymbol{R}_t)_+$

Table 1.3 – ML-Poly algorithm, at time t after y_t is given. The vectors $\boldsymbol{\eta_t}$ and $\boldsymbol{R}_t$ have M coordinates, respectively $\eta_{m,t}$ and $R_{m,t}$. The notation $(\cdot)_+ = \max(\cdot, 0)$ is used.

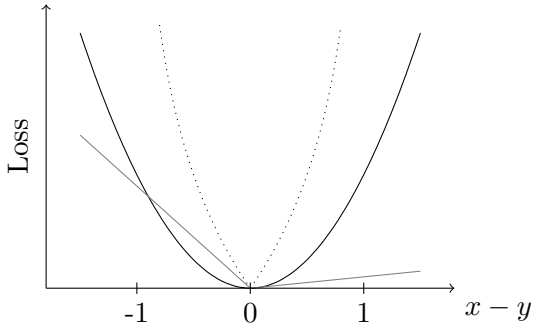


Figure 1.8 – Deterministic losses: quadratic loss (black), quantile loss of level 0.9 (gray), log loss for $y \in \{0, 1\}$ (dotted black).

As opposed to the aforementioned regret bounds, the bound of ML-Poly is of second-order due to the term $\sum_{t=1}^{T} (\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2$. The interested reader is referred to Gaillard et al. [GSE14] for a detailed analysis of second-order bounds, and to Koolen and Van Erven [KV15], Luo and Schapire [LS15], and Wintenberger [Win17] for further examples of algorithms showing second-order bounds. These bounds are adaptive in the sense that the algorithm performs well in the worst-case, but shows improved guarantees for easy data. The worst case scenario gives a bound $\mathcal{O}(\sqrt{MT \ln T})$, indicating that even in the worst case, the weighted forecast will perform at least as well as the best forecast. Besides, in the case of i.i.d. sequences of losses, with one expert significantly better than the others, the regret bound is practically constant.

1.2.4 Examples of loss functions

In the most common approach, the forecaster provides a point forecast ($\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$). Noting $s(\boldsymbol{u}, \boldsymbol{x}) = \widehat{y}$, the following losses are standard examples:

— Square loss: $(\widehat{y} - y)^2$.

— Quantile loss of level α: $(\mathrm{H}(\widehat{y} - y) - \alpha)\,(\widehat{y} - y)$, where H is the unit step function ($\mathrm{H}(x) = 1$ if $x > 0$ and zero otherwise). Let y be a random variable described by the CDF F, the quantile loss of level α is minimized in average over the distribution of y by $\mathrm{F}^{-1}(\alpha)$. The absolute loss $|s(\boldsymbol{u}, \boldsymbol{x}) - y|$ equals twice the quantile loss of level 0.5.

— Log loss: $-y \ln(\widehat{y}) - (1 - y) \ln(1 - \widehat{y})$; often used in classification with $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = [0, 1]$ or $\mathcal{Y} = \{0, 1\}$.

See Figure 1.8 for a graphical representation of the losses.

In a probabilistic forecasting, the probability density function (PDF) g and the related CDF G are delivered, while the value $y \in \mathbb{R}$ is observed. For linear models $\mathrm{G} = \sum_m u_m \mathrm{G}_m$, the forecaster delivers a so-called linear opinion pool. In this setting, standard examples are

— Quadratic score: $-2g(y) + \int g^2$.

— Logarithm loss: $-\log(g(y))$.

— Hyvärinen score: $2\left(\frac{g''(y)}{g(y)}\right) - \left(\frac{g'(y)}{g(y)}\right)^2$.

— Continuous ranked probability score (CRPS): $\int(\mathrm{G}(\theta) - \mathrm{H}(\theta - y))^2 \mathrm{d}\theta$.

In our contributions, we use the square loss in Chapter 2 to improve deterministic forecasts of solar radiations, and the CRPS in Chapter 3 from a theoretical perspective in online learning and in Chapter 5 for practical applications with PV forecasts. The last section of this introduction is dedicated to a specific type of losses for probabilistic forecasting, which includes the CRPS.

1.3 Probabilistic forecasting with non-local strictly proper scoring rules

In this section, we introduce non local strictly proper scoring rules and explain why we resort to them.

Strict propriety. We are interested in a strictly proper scoring rule in the context of probabilistic forecasting, see Gneiting and Raftery [GR07]. Say a forecaster provides a probabilistic forecast to predict y. We consider that y is a random variable described by the CDF F. The quality of the forecast is measured by a score $S(\mathrm{G}, y)$, where G is a CDF describing the probabilistic forecast. The scoring rule is said to be proper if $\mathrm{E}_y[S(\mathrm{F}, y)] \leqslant \mathrm{E}_y[S(\mathrm{G}, y)]$ and strictly proper if the strict inequality holds. In other words, with a strictly proper scoring rule, the average score is minimized only by the correct distribution F.

Locality. Let $\{A_1, \ldots, A_K\}$ be a partition of K possible events. By definition, two events A_i and A_j cannot occur simultaneously. The forecaster delivers $\boldsymbol{p} = (p_1, \ldots, p_K)$ reflecting an opinion on the probability of occurrence of each event. If the event A_i occurred, a local scoring rule of $\boldsymbol{p}$ depends only on p_i . In other words, a scoring rule is said to be local if the score only depends on the predicted probability of the event that actually occurred. Locality of the scoring rule is not always a desired property [BS07b]. The logarithm score was shown to be a natural choice of strictly proper local scoring rule in the non-binary case ($K > 2$), see Bernardo [Ber79] and Benedetti [Ben10] for example.

Why do we use non local strictly proper scoring rules ? Our objective is to provide probabilistic forecasts. To do so, we work with an ensemble of forecasts, which is conveniently described by a CDF. The corresponding probability density function is a combination of Dirac distributions, reaching 0 almost everywhere. The events "y is between θ_k and $\theta_k + \mathrm{d}\theta_k$" are therefore poorly predicted, and local scoring rules are hardly applicable in our case.

We introduce the case of binary events in Section 1.3.1, and probabilistic scores for continuous variables in Section 1.3.2. We follow here Shuford et al. [SAE66], Schervish [Sch89], and Buja et al. [BSS05] to give a simple explanation on the construction (but not the choice) of non-local strictly proper scoring rules. A discussion on the choice of strictly proper scoring rule may be found in Merkle and Steyvers [MS13] and Lerch et al. [Ler+15]. A thorough analysis of non local strictly proper scoring rules and the decomposition with quantile and thresholds is given by Ehm et al. [Ehm+16]. A simple example is given in Section 1.3.3, to provide a better understanding of the evaluation of a probabilistic forecast by the celebrated CRPS.

1.3.1 Binary case

We start from the binary case, where the forecaster wishes to predict the output of a single event and delivers the probability p of occurence of the event according to his opinion. A scoring rule ℓ for binary event is local and of the form

$$\ell(p) = \mathbf{1}_{ev} S_1(p) + (1 - \mathbf{1}_{ev}) S_0(1-p)\,,$$

where $\mathbf{1}_{ev} = 1$ if the event occurs and 0 otherwise. The forecaster suffers $S_1(p)$ when the event occurs and $S_0(1-p)$ when the event does not occur.

The strict propriety imposes that the average score $\mathrm{E}(\ell(p))$ is minimized if (and only if) p is the true probability of occurrence f, and accordingly:

$$f S_1'(f) - (1-f) S_0'(1-f) = 0\,, \tag{1.1}$$

thanks to the stationnarity condition $\frac{\mathrm{d}\ell}{\mathrm{d}p}(p=f) = 0$. Equivalence between the stationnarity condition and strict propriety is demonstrated by Shuford et al. [SAE66].

The scoring rules may be taylored with respect to cost functions, and an infinite number of scoring rules are strictly proper. We restrict the study to the symmetrical case $S_0 = S_1$, where we have

$$f S_1'(f) = (1-f) S_1'(1-f)\,,$$

according to Equation 1.1. We define $\zeta(p) = p S_1'(p)$, with ζ being symmetrical with respect to 1/2: $\zeta(p) = \zeta(1-p)$. The strict propriety necessitates that $\zeta(p)$ is non-zero for $p \neq f$ because for $p \in]0,1[$:

$$\frac{\mathrm{d}\,\mathrm{E}(\ell(p))}{\mathrm{d}p} = \frac{f}{p}\zeta(p) - \frac{1-f}{1-p}\zeta(1-p) = \frac{\zeta(p)}{p(1-p)}(f-p)\,.$$

Losses are usually negatively oriented, meaning that $\ell(p)$ should be as low as possible. Since the loss should decrease when p gets closer to f, we have $\zeta(p) < 0$.

This writing of the loss can also be interpreted with quantile losses. Let α be a decision threshold. We consider the decision-thresholded loss where a false negative costs $1-\alpha$ (for $\mathbf{1}_{ev} = 1$ and $p \leqslant \alpha$), and a false positive costs α (for $\mathbf{1}_{ev} = 0$ and $\alpha < p$). The total loss $\ell(p)$ is a weighted sum of decision-thresholded losses, where the importance of the decision thresholds are given by $\omega(\alpha)$:

$$\ell(p) = \mathbf{1}_{ev} \int_p^1 \omega(\alpha)(1-\alpha)\mathrm{d}\alpha + (1-\mathbf{1}_{ev}) \int_0^p \omega(\alpha)\alpha \mathrm{d}\alpha\,.$$

$\omega(p)$	$\ell(p)$
$(p(1-p))^{-3/2}$	$\mathbf{1}_{ev}\sqrt{\frac{1-p}{p}} + (1-\mathbf{1}_{ev})\sqrt{\frac{p}{1-p}}$
$(p(1-p))^{-1/2}$	$\mathbf{1}_{ev}\arcsin(\sqrt{1-p}) + (1-\mathbf{1}_{ev})\arcsin(\sqrt{p}) - \sqrt{p(1-p)}$
$p(1-p)$	$\mathbf{1}_{ev}(\frac{(1-p)^3}{3} - \frac{(1-p)^4}{4}) + (1-\mathbf{1}_{ev})(\frac{p^3}{3} - \frac{p^4}{4})$

Table 1.4 – Example of scoring rules in the beta family $\omega(p) = p^{a-1}(1-p)^{b-1}$.

We identify $S_1(p) = \int_p^1 \omega(\alpha)(1-\alpha)\mathrm{d}\alpha$ and $S_0(1-p) = \int_0^p \omega(\alpha)\alpha\mathrm{d}\alpha$. When $\omega(\alpha)$ is symmetrical with respect to $1/2$ ($S_1 = S_0$), we find $\zeta(p) = pS_1'(p) = -\omega(p)p(1-p)$. We see that the function ζ is closely related to the weighting function ω of the thresholds. For any strictly proper score defined by S_1, it is possible to define $\omega(p) = -S_1'(p)/(1-p)$, and consider the loss as a weighted sum of decision-thresholded losses.

Two classical examples are the quadratic (Brier) score and the logarithm score:

$$\ell_{sq}(p) = \mathbf{1}_{ev}(1-p)^2 + (1-\mathbf{1}_{ev})p^2 = (\mathbf{1}_{ev} - p)^2$$
$$\ell_{ln}(p) = -(\mathbf{1}_{ev}\ln(p) + (1-\mathbf{1}_{ev})\ln(1-p))$$

respectively obtained for $\omega(p) = 2$ and $\omega(p) = \frac{1}{p(1-p)}$. Other examples from Buja et al. [BSS05] are summarized in Table 1.4.

1.3.2 Ranked and continuous case

We now consider events of the type: "y is below the threshold θ_k". With K threshold-type events, the binary case of Section 1.3.1 easily extends to

$$\ell(\boldsymbol{p}, y) = \sum_{k=1}^{K} \phi_k \left[\mathrm{H}(\theta_k - y)S_1(p_k) + \mathrm{H}(y-\theta_k)S_0(1-p_k)\right] ,$$

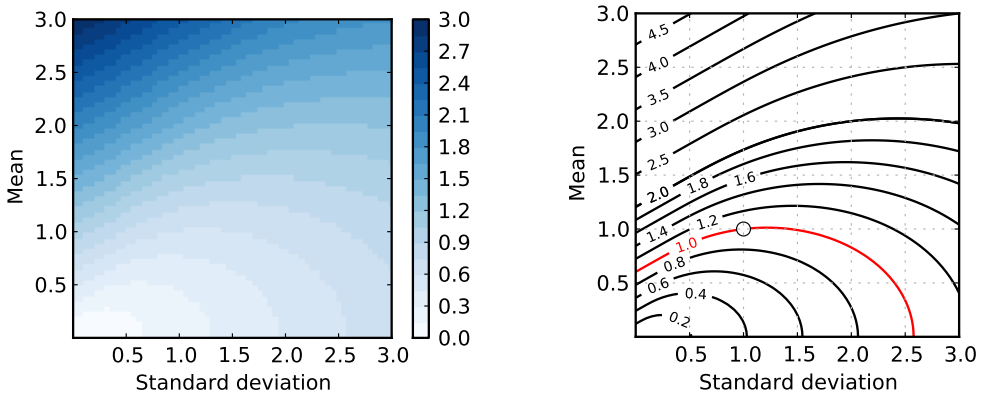
where $\boldsymbol{p}$ is the list of predicted probabilities p_k for the events, the weights $\phi_k > 0$ define the importance of the thresholds θ_k, and H is the Heaviside step function. We adopt the bold notation $\boldsymbol{p}$ to describe the list or vector with coordinates p_k. The events "y is below the threshold θ_k" are not a partition of events, consequently this scoring rule is not local. If the dimension of the weights ϕ_k and the dimension of the observation y coincide, the loss $\ell(\boldsymbol{p}, y)$ has the same dimension as y.

In the continous case, for each threshold θ, the weighting ϕ_k is related to $\phi(\theta)\mathrm{d}\theta$. The following continous score is written with the CDF notation:

$$\ell(\mathrm{G}, y) = \int \phi(\theta) \left[\mathrm{H}(\theta - y)S_1(\mathrm{G}(\theta)) + \mathrm{H}(y-\theta)S_0(1-\mathrm{G}(\theta))\right] \mathrm{d}\theta \,.$$

Using a uniform weighting scheme over the decision thresholds $\phi(x) = 1$ with the quantile weighting $\omega(p) = 2$ and $\omega(p) = \frac{1}{p(1-p)}$, we find the CRPS and the Continuous Ranked Ignorance score (CRIGN) defined by:

$$\mathrm{CRIGN} = -\int \mathrm{H}_y \ln \mathrm{G} + (1-\mathrm{H}_y)\ln(1-\mathrm{G})$$
$$\mathrm{CRPS} = \int \mathrm{H}_y(1-\mathrm{G})^2 + (1-\mathrm{H}_y)\mathrm{G}^2 = \int (\mathrm{H}_y - \mathrm{G})^2 \,,$$



(a) Map of CRPS of normal distributions for the observation $y = 0$, depending on the mean and standard deviation of the distribution.

(b) Ratio of CRPS of normal distribution defined by its mean and standard deviation against the CRPS of $\mathcal{N}(1, 1)$.

Figure 1.9 – CRPS for normal distribution defined by its mean and standard deviation, for the observation $y = 0$.

with H_y the Heaviside function centered on y. At last, the CRPS and the CRIGN are two examples of non local strictly proper scoring rules. By construction, they enjoy desirable properties of being strictly proper, compatible with ensemble forecasts, and show decomposition properties [Her00; TA12].

1.3.3 Examples with the CRPS

In this section, we provide simple examples with the CRPS, as an evaluation tool. The idea is that for a fixed value of the observation y, several CDFs may reach the same score. We illustrate how these CDFs are related to each other.

In Figure 1.9(a), we show the map of the CRPS for CDF of normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2, for the observation $y = 0$. For a fixed value of the standard deviation σ, we see that the lowest value of the CRPS is reached for $\mu = y = 0$. Besides, for a fixed value of the mean μ, the lowest value of the CRPS is not reached for the lowest spread $\sigma = 0$. Instead, the lowest CRPS value is reached when the spread σ is slightly higher than the error $|\mu - y|$ on the mean. The idea behind is that, for an incorrect forecast with an error on the mean, the spread of the distribution lowers the effect of the error $|\mu - y| > 0$.

We compare CRPS values with the case $\mathcal{N}(1, 1)$ in Figure 1.9(b). The same CRPS values are obtained for $g_a = \mathcal{N}(0.6022, 0)$, $g_b = \mathcal{N}(1, 1)$ and $g_c = \mathcal{N}(0, (2.577)^2)$, with respective CDFs G_a, G_b and G_c. The distribution g_a has no spread and a small error of 0.6 on the mean, while the distribution g_c has a high spread but no error on the mean. We give an illustration of the CDFs G_a, G_b and G_c, the PDFs g_a, g_b and g_c and the errors $(\mathrm{G}_a - \mathrm{H}_y)^2$, $(\mathrm{G}_b - \mathrm{H}_y)^2$, $(\mathrm{G}_c - \mathrm{H}_y)^2$ in Figure 1.10. The strong effect of the square in the CRPS is quite noticeable on the representation of $(\mathrm{G}_b - \mathrm{H}_y)^2$, and more generally on the domains where $|\mathrm{G}(\theta) - \mathrm{H}(\theta - y)| \ll 1$. Consequently, the CRPS is not

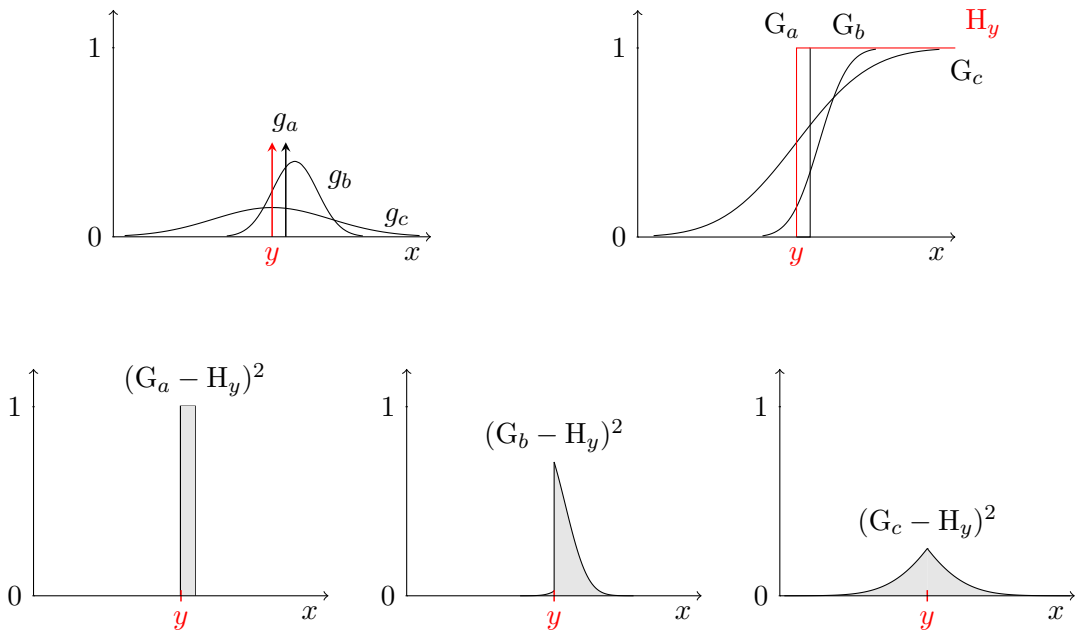


Figure 1.10 – Representation of the PDFs $g_a = \mathcal{N}(0.6022, 0)$, $g_b = \mathcal{N}(1, 1)$ and $g_c = \mathcal{N}(0, (2.577)^2)$ (top left), and their respective CDFs G_a, G_b and G_c (top right), reaching the same CRPS of 0.602 at 0.001 precision (for $y = 0$). The squared difference between H_y the Heaviside function centered on y and respectively G_a, G_b and G_c are shown at the bottom.

very sensitive to the distribution tails. On this subject, we address in Chapter 4 the question of other scoring rules than the CRPS.

Thesis outline

In Chapter 2, we study multiple ensembles of forecasts of solar radiation, and use sequential aggregation to improve forecast maps of solar radiation.

In Chapter 3, we study the CRPS in the context of model mixture, specifically the bias of the CRPS of an ensemble of forecasts. We introduce sequential aggregation with the CRPS. In Chapter 4, we extend the results of Chapter 3 to other losses than the CRPS admitting a threshold or a quantile decomposition. We also investigate the question of noisy observations.

In Chapter 5 and 6, case studies of probabilistic forecasts of PV power are provided. We show that the methods introduced in Chapter 3 improve the quality of the forecasts for both deterministic and probabilistic evaluation tools. Chapter 7 focuses on PV power forecasts with intraday updates for insular systems such as Réunion and Corsica.

Publications

Chapter 2 is based on the paper Thorey et al. [Tho+15]:

Thorey, J., Mallet, V., Chaussin, C., Descamps, L. and Blanc, P. (2015). Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database. Solar Energy, 120, 232-243.

Chapter 3 is based on the paper Thorey et al. [TMB16]:

Thorey, J., Mallet, V. and Baudin, P. (2017), Online learning with the Continuous Ranked Probability Score for ensemble forecasting. Q.J.R. Meteorol. Soc., 143: 521–529.

Chapter 5 is submitted as a research article to the *International Journal of Forecasting*.

2 Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database

Medium-range forecasts (one day to two weeks) of solar radiation are commonly assessed with a single forecast at a given location. In this paper, we forecast maps of surface solar irradiance, using ensembles of forecasts from the THORPEX Interactive Grand Global Ensemble (TIGGE) with a 6-h timestep. We compare our forecasts with observations derived from MeteoSat Second Generation (MSG) and provided by the HelioClim-3 database as gridded observations over metropolitan France. First, we study the ensembles from six meteorological centers. Second, we use sequential aggregation to linearly combine all the forecasts with weights that vary in space and time. Sequential aggregation updates the weights before any forecast, using available observations. We use the global numerical weather prediction from the European Center for Medium-range Weather Forecasts (ECMWF) as a reference forecast. The issue of spatial resolution is discussed because the low resolution forecasts from TIGGE are compared to high resolution irradiance estimated from MSG data. We found that the TIGGE ensembles are under-dispersed but rather different from one to another. Aggregation decreases the forecast error by 20%, and produces a more realistic spatial pattern of predicted irradiance.

This chapter was published as the paper: Thorey, J., Mallet, V., Chaussin, C., Descamps, L., and Blanc, P. « Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database ». In: *Solar Energy* 120 (Oct. 2015), pp. 232–243.

Contents

2.1 Introduction

Solar radiation forecasts and especially global horizontal irradiance (GHI) forecasts are needed for the integration of photovoltaic power (PV). The increasing installed capacity of photovoltaic power requires that solar radiation forecasts be always more accurate in terms of spatial and temporal resolutions.

Many meteorological centers provide solar radiation forecasts with two strategies: either a single deterministic forecast, or ensemble forecasts generally with coarser resolution. Deterministic forecasts have been extensively studied for solar and photovoltaic forecasts. Inman et al. [IPC13] and Espinar et al. [Esp+10] review numerous modeling techniques to generate solar and PV forecasts from meteorological variables. Deterministic predictions from various meteorological centers are compared by Lorenz et al. [Lor+09a] and Perez et al. [Per+13] in a broad range of sites. In the previously cited papers, some forecasting techniques resort to combinations of forecasts. These combinations are derived from a regression over a moving time-window and are applied to a few members.

Vernay et al. [V+13] listed several available maps of solar radiation deriving from cloud products of satellite observations. Even though solar radiation forecasts cover large areas, they are usually compared to ground measurements (PV or solar) or to satellite observations only at measurement sites [e.g., GDM80, among the firsts]. Indeed maps of satellite observations are not broadly used to assess the accuracy of solar radiation forecasts. Morcrette [Mor91] used satellite observations as reference to assess the performance of numerical weather predictions (NWP), but not predictions of solar irradiance (e.g., short-wave radiation). Perez et al. [PSZ97] studied the interactions between satellite observations and measurement sites with respect to the distances between sites. Due to the variety of error causes, Thelen and Edwards [TE13] restricted the comparison between NWP and satellite observations to reflectance for short-wave radiation. Dehghan et al. [Deh+14] give emphasis on the spatial resolution of both NWP and satellite data at ground measurement sites.

Ensemble forecasts are classical in meteorology for any field with large uncertainty and for uncertainty quantification. However, no article using ensemble forecast for solar radiation was found in the literature, despite several conference presentations. Still, Yokohata et al. [Yok+12] studied climatological ensembles of top atmosphere radiation and radiation in cloud-free conditions.

In the framework of sequential aggregation, a single forecast is built as a linear combination of the ensemble members. The weights of the combination may depend on both time and space. The resulting aggregated forecast is hopefully more skillful than the ensemble members. Cesa-Bianchi and Lugosi [CL06] detail the strong mathematical background of these methods, which is summarized and tested by Stoltz [Sto10] and Mallet et al. [MSM09] and Mallet [Mal10] on forecasts of respectively electricity consumption and ozone concentrations.

We propose here to compare the ensemble forecasts of solar radiation from TIGGE (THORPEX Interactive Grand Global Ensemble [Bou+10]) and to combine them. The satellite observations from HelioClim are used in this article as reference observations. The use of both ensemble forecasts (from several sources) and satellite observations

makes it possible to generate an aggregated forecast with local combinations on the spatial grid. The Integrated Forecast System (IFS) from ECMWF produces our reference forecast.

In Section 2.2, we describe the TIGGE data sets and we study the performance of the TIGGE ensembles. The HelioClim satellite observations are introduced in Section 2.2.4, along with a detailed comparison with TIGGE forecasts. Our sequential-aggregation strategy is introduced in Section 2.3. It is applied in Section 2.4, where the analysis of the results includes the composition of the ensemble, the spatial patterns in the forecast maps, the forecast time horizon and the sensitivity to the aggregation parameters.

2.2 Analysis of TIGGE solar radiation and HelioClim database

2.2.1 Description of TIGGE data

Several meteorological centers provide free-of-charge ensemble forecasts of solar radiation in TIGGE (Table 2.1). The data sets are available on TIGGE with a 2-day delay after the model ran. Detailed studies of the ensemble forecasts from TIGGE were achieved on geopotential height [Bui+05], and 850-hPa and 2-m temperatures [Hag+12] for example. As far as we know, no such study exists for solar radiation.

The temporal resolution of TIGGE forecasts is 6 h with time horizon up to 15 days. Most meteorological centers provide at least two forecast sets per day. For the sake of clarity and because of the 6-h timestep of TIGGE forecasts, we focus on one forecast set per day for each meteorological center. With these constraints, 6 meteorological centers provide an ensemble: China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), MetOffice (UKMO), Korea Meteorological Administration (KMA), Centro de Previsao Tempo e Estudos Climaticos (CPTEC), and Météo-France (M.-F.). We name "whole ensemble" the ensemble including all 158 members without consideration of their origins, as opposed to the 6 TIGGE center ensembles.

While each center ensemble has a native spatial resolution, our TIGGE data sets are obtained on a common regular grid with the spatial resolution of $0.25° \times 0.25°$, which is finer or close to the native resolutions. The resolution of the TIGGE data sets is coarser than the $0.125° \times 0.125°$ resolution of the ECMWF deterministic forecast. Our study focuses on Metropolitan France and the surrounding areas for the 06:00–12:00 UTC accumulation period of day D with model runs starting at 18:00 or 24:00 UTC in day D-1. Thus we study the forecasts for either 12 h or 18 h of lead time. We do not study daily radiation but focus on the shortest timestep available in TIGGE. The flux values from TIGGE are averaged over the 6-h timestep so that the values of the forecasts for 12:00 UTC are expressed in $\mathrm{W\,m^{-2}}$ and correspond to the averaged flux between 06:00 and 12:00 UTC.

The nature of the solar radiation data sets in TIGGE is mostly net shortwave solar radiation (SSR) as defined for classical meteorological fields. Net shortwave solar radiation is the fraction of the downwards shortwave solar radiation (SSRD) absorbed by the ground on an horizontal plane. Note that KMA data are different from the other

Center	Origin	Number of members	Run
CMA	China	14	24:00
ECMWF	UE	50	24:00
UKMO	UK	23	24:00
KMA	Korea	23	24:00
CPTEC	Brazil	14	24:00
Météo-France	France	34	18:00

Table 2.1 – Overview of TIGGE ensembles available for solar radiation, with forecasts starting from 18:00 or 24:00 UTC for the following 108 to 360 hours (horizon). In total, there are 158 ensemble members.

TIGGE data sets and are SSRD data. In the context of photovoltaic production, we are interested in SSRD. The well-known global horizontal irradiance (GHI) refers to SSRD. The albedo coefficient α is the reflection coefficient of the ground. The albedo defines a relationship between SSR and SSRD, where the incident flux is divided between the absorbed flux and the reflected flux. Thus we deduce that:

$$\text{SSRD} = \frac{\text{SSR}}{1-\alpha}\,. \tag{2.1}$$

Depending on the ground surface, the albedo coefficient can vary in space and time.

2.2.2 Analysis of the TIGGE ensembles of forecasts

Now we analyze and compare the ensembles of forecasts over the area spanning 41° to 51.50° in latitude and −5.50° to 10° in longitude. Two data sets are used in our study. The main data set consists of 350 consecutive days starting on 2012-01-02. The secondary data set is dated from 2011-06-09 to 2011-09-05 due to data availability and includes only 100 random locations. The secondary data set is only used as learning data set. An example of the annual average of the 6-hour forecasts at 12:00 UTC for one center (KMA) is provided in Figure 2.1, in order to show the large spatial variability of the average forecast.

We propose two sorting procedures so as to consistently number the forecasts in time. This step is required for the weights in our aggregation to be clearly associated with a given ensemble member and therefore to be meaningful. The two sorting procedures rely on the rank at each grid point. At any time and location, the first member of a sorted ensemble always provides the lowest value, and the last member gives the highest value. The first sorting procedure is applied to the 158 members of the whole ensemble, and the second sorting procedure is applied separately in each center ensemble.

Correlation coefficients between members are computed in order to quantify the similarities between the members (Figure 2.2). In the correlation matrices, the rectangles and squares materialize some separation between the different ensembles. Indeed the correlation matrices reveal that the members are especially close to one another within the same ensemble. In other words, one member from a given meteorological center ensemble is more correlated to another member of the same ensemble than to another

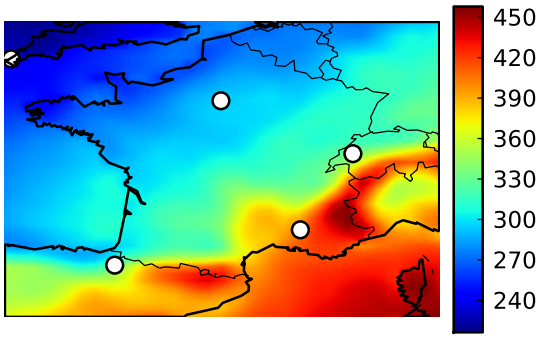


Figure 2.1 – Annual average of 6-hour forecasts at 12:00 UTC in W m^{-2}, for the KMA ensemble mean in year 2012. The white circles exhibit the locations of ground observation sites for HelioClim evaluation.

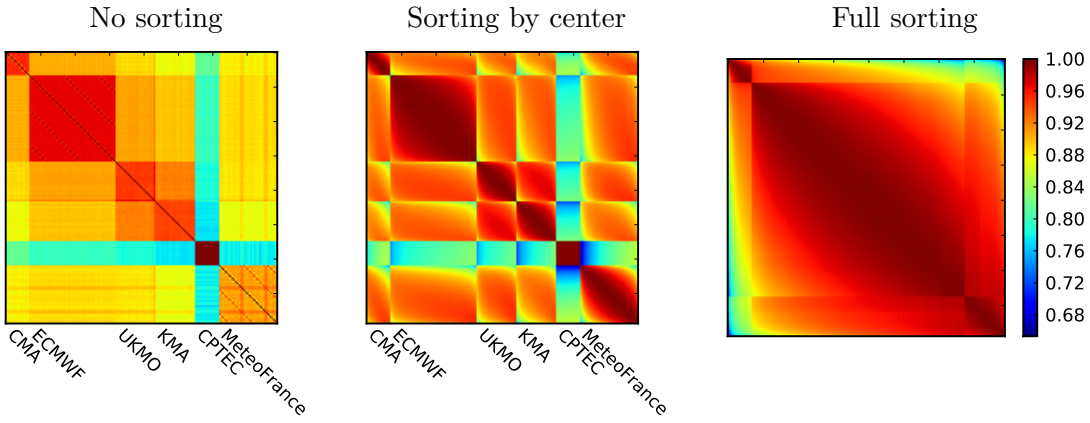


Figure 2.2 – Matrices of correlation between members (R^2 matrices). Each row and column of the correlation matrices is dedicated to one given member. The rows and columns appear in the same order. Each matrix entry is the correlation coefficient between the merged data (all timesteps, 100 random locations) of a pair of members. On the left, the correlations are shown between raw members, as they are retrieved from TIGGE. In the middle, the correlations are computed after sorting within each center ensemble. On the right, the correlations are obtained after a full sorting of the forecasts.

member from a different ensemble. The CPTEC ensemble is very distinguishable due to the extremely high correlations between its own members and also due to the low resemblance between its members and the others.

Sorting has generally two effects on the ensembles. First, the sorted members with close ranks have higher correlations among them than non-sorted members. Second, the pairs of members with the lowest correlation coefficient are found between sorted members of extreme rank.

2.2.3 Reference performance measures

The strengths and weaknesses of the statistical indicators commonly used in solar forecasting are developed in Hoff et al. [Hof+13]. The well-known root mean square error (RMSE) and mean absolute error (MAE) are classical performance indicators. The RMSE of the predictions $\widehat{y}$ with respect to the observations y over the set $\mathcal{S}$ is given by

$$RMSE = \sqrt{\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}(\widehat{y}_s - y_s)^2}\,, \tag{2.2}$$

where $|\mathcal{S}|$ is the number of elements (cardinality) in $\mathcal{S}$, and s indexes space or time or both. In case s describes all locations at one single timestep, the spatial RMSE is computed. The MAE is calculated in a similar way as

$$MAE = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}|\widehat{y}_s - y_s|\,. \tag{2.3}$$

It is noticeable that the errors $\widehat{y}_s - y_s$ are computed independently for each s, hence errors due to geographical or temporal shifts are not detected as such. In case of missing predictions, the missing indices are excluded from the set $\mathcal{S}$.

The average observed value is introduced to define relative indicators; e.g. for the relative RMSE:

$$rRMSE = \frac{RMSE}{\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} y_s}\,, \tag{2.4}$$

where rRMSE and RMSE are computed over the same set $\mathcal{S}$. Except for the scores of Table 2.2, HelioClim is used as the observation y_s. Because our main interest is to provide information over land and not over sea, all scores and regression coefficients are given for land locations only.

2.2.4 Comparison with HelioClim

Description of HelioClim data

Provided by Transvalor (Armines), HelioClim satellite observations are based on MSG satellite data (Meteosat Second Generation). The operational version HC3-v4 is used here. The HelioSat2 method [RLW04] has been deployed to generate this HelioClim database [Bla+11]. The satellite estimations rely on instantaneous reflectance

Station	Bias	RMSE	MAE
Camborne	-1.4 % (-3.3)	10.6 % (25.6)	8.3 % (20.0)
Palaiseau	3.4 % (8.5)	10.7 % (26.9)	8.6 % (21.5)
Payerne	-6.9 % (-18.3)	14.4 % (38.3)	10.9 % (29.1)
Carpentras	1.3 % (4.6)	8.3 % (28.9)	6.7 % (23.3)
Cener	2.9 % (9.7)	9.5 % (32.0)	7.6 % (25.6)

Table 2.2 – Evaluation of HelioClim-3 estimation of SSRD compared to in-situ measurements (daily average of daytime data). The relative scores are given, followed by the absolute scores in brackets (in $\mathrm{W\,m^{-2}}$).

measurements. Data are acquired every 15 min, converted to an estimation of the incident radiation flux, and averaged over the 15-min timestep. The spatial resolution of HelioClim over France is natively between 3 and 5 kilometers; our HelioClim data were retrieved with a spatial resolution of $1/12°$, which is already much finer that the resolution of the forecasts.

Our zone of interest includes five BSRN stations (Baseline Surface Radiation Network) [Ohm+98] for the evaluation of HelioClim performance. The stations are located in Camborne (United Kingdom), Cener (Spain), Carpentras (South France), Palaiseau (North France), and Payerne (Switzerland), as exhibited in Figure 2.1. The evaluation results (Table 2.2) are computed over several years and show an average relative RMSE of 10.7% and an average relative MAE of 8%.

In Section 2.4, we wish to produce our own forecasts based on the previously described ensembles and we wish our forecasts to be at the finest available resolution, which is the spatial resolution of our HelioClim data. Consequently our study is carried out at the resolution of HelioClim. All the forecasts are interpolated by bilinear interpolation to reach the same resolution of $1/12°$, for a total amount of 127×187 grid points. On the high-resolution grid, the forecasts vary spatially slowly compared to the satellite observations.

TIGGE ensembles and HelioClim

The difference of nature between the HelioClim incident radiation (SSRD) and the ground-absorbed radiation from TIGGE (SSR) prevents a direct comparison. Therefore we test several conversion methods in Appendix 2.A. The method of linear conversion (lin) based on historical data shows the best RMSE. This method is used until the end of the section.

The monthly RMSEs are impacted by the seasonal variability (Figure 2.3). Solar radiation forecasting is more difficult between April and July, during the brightest days with large variability. It is worthy of notice that the ranking of the ensemble means is steady over time. The ensemble means of KMA and UKMO (after linear conversion) show better scores than the reference forecast, while their spatial resolution is natively poorer. An improved forecast, called TIGGE-mean, may be built as the mean of the TIGGE ensemble means (CPTEC excluded) with linear conversion. The score of the TIGGE-mean lies at $66.9\,\mathrm{W\,m^{-2}}$, which is better than any TIGGE ensemble mean.

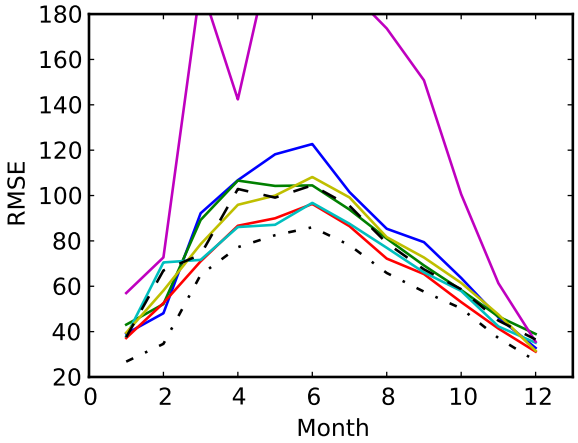


Figure 2.3 – Monthly RMSEs of TIGGE ensemble means in $\mathrm{W\,m^{-2}}$, after linear conversion (lin); (blue: CMA), (green: ECMWF), (red: UKMO), (cyan: KMA), (magenta: CPTEC) and (yellow: Météo-France). The dashed line is the score of the reference forecast. The dashed-dotted line is the score of a typical aggregated forecast (see Section 2.4.2).

The same analysis was undertaken with the MAE (not shown) and reveals similar trends: $54.4\,\mathrm{W\,m^{-2}}$ (reference forecast MAE), $50.2\,\mathrm{W\,m^{-2}}$ (UKMO ensemble mean with linear conversion MAE), $48.0\,\mathrm{W\,m^{-2}}$ (TIGGE-mean MAE).

We highlight the fact that no center ensemble is steadily the closest to the satellite observations, whatever the score discrepancy described above. If we track over the consecutive timesteps the origin of the best member at each location, we find that this origin changes from one timestep to the next with a frequency of 75%. Furthermore, the proportion of times and locations where the best member belongs to a given center is reported here: reference ECMWF 3%, CMA 10%, ensemble ECMWF 26%, UKMO 19%, KMA 18%, CPTEC 2%, Météo-France 23%. Considering the fact that the ECMWF reference forecast is one single member compared to 158 members, its frequency of being the closest member to the observation is rather high. On the opposite, the CPTEC ensemble and, to a lesser extent, the CMA ensemble show the lowest frequencies.

Rank histograms [And96; TVS99; HC97] evaluate the quality of the spread of an ensemble. Each observation is given a rank, which corresponds to the number of members with lower values than the considered observation. Then the distribution of rank frequencies can reveal the presence of under-dispersion (U-shaped histogram), over-dispersion, and biases. The rank histograms of the center ensembles and the rank histogram of the whole 158-member ensemble are shown in Figure 2.4 and 2.5, for data converted by center (lin). The values of the bars of each histogram are normalized with respect to the total number of observations, so that the height of the bars of the ideally flat histogram is always 1. There are clear outliers on all histograms, which is a clear sign of general under-dispersion [Ham01]. The frequencies with which observations fall within an ensemble envelop are: CMA 34%, ECMWF 45%, UKMO 58%, KMA 56%, CPTEC 1%, Météo-France 59%, whole ensemble 89%.

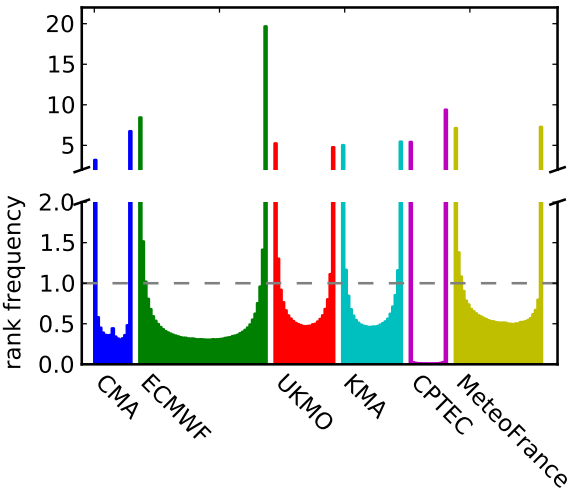


Figure 2.4 – Rank histograms of the center ensembles. In the ideal case of a flat histogram, the height of all the bars would be equal to 1.

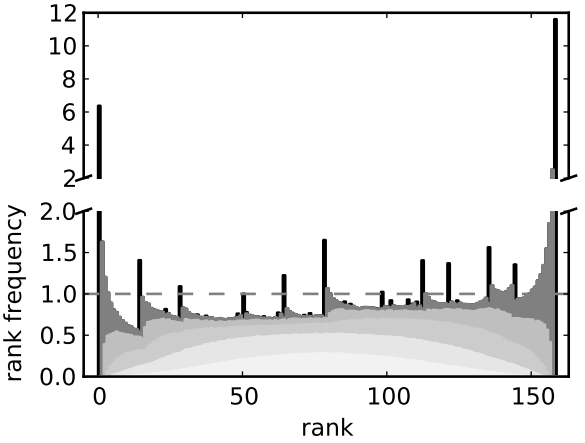


Figure 2.5 – Rank histograms for the whole ensemble. In the ideal case of a flat histogram, the height of all the bars would be equal to 1. The gray scale indexes the number of ensembles whose spread contains the observation, from black (no ensemble) to white (all ensembles, but not observed because of the very low dispersion of CPTEC).

The whole ensemble is less under-dispersed than each center ensemble and exhibits a rank structure with peaks. The ranks of the major inner peaks (14, 28, 37, 78, 112, 121, 135 and 144) correspond to combinations of ensemble sizes, starting from either extreme peak. For example, the peak at rank 28 corresponds to the combination of the ensemble sizes of CPTEC and CMA (both 14). The peak at rank 135 corresponds to an ensemble of 23 members (UKMO or KMA) starting from the outer peak at rank 158. The peaks are generated by the 4% share of the observations (inner black bars) which do not fall within any of the center ensemble envelops. Indeed these observations are necessarily indexed at ranks related to combinations of ensemble sizes. We conclude that the peaks are due to the variety of the under-dispersed ensembles.

2.3 Ensemble forecast strategy: sequential aggregation

In this section we detail the principles of sequential aggregation. In particular, we explain the method of discounted ridge regression. The performance and robustness of discounted ridge regression have been tested for the case of air quality [MSM09]. While time series of scalar fields were considered above, in this section we only consider scalar time series.

2.3.1 Notation

Let $x_{m,t}$ describe the m-th member of our forecast ensemble at time t, with $m \in \{1, \ldots, M\}$ indexing the members and $t \in \{1, \ldots, T\}$ indexing the forecast timesteps. The vector $\boldsymbol{x}_t$ refers to the ensemble of forecasts $[x_{1,t}, x_{2,t}, \ldots, x_{M,t}]^\top$. At each timestep, the members should be conveniently combined with the weights $w_{m,t}$ to generate the forecast

$$\widehat{y}_t = \sum_{m=1}^{M} w_{m,t} x_{m,t} \tag{2.5}$$

of the observation y_t.

2.3.2 Sequential aggregation: method

The aggregation weights $w_{m,t}$ are updated before the forecast step t, using only past observations y_1, y_2, …, y_{t-1} and past simulations $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_{t-1}$. For the discounted ridge regression with parameters (λ, γ), the weight vector $\boldsymbol{w}_t = [w_{1,t}, w_{2,t}, ..., w_{M,t}]^\top$ is found through minimization of

$$J(\boldsymbol{u}) = \lambda \|\boldsymbol{u} - \boldsymbol{w}_{\text{ref}}\|_2^2 + \sum_{t'=1}^{t-1} \beta_\gamma(t - t') \times (y_{t'} - \boldsymbol{u} \cdot \boldsymbol{x}_{t'})^2 \tag{2.6}$$

with

$$\beta_\gamma(t - t') = 1 + \frac{\gamma}{(t - t')^2} \tag{2.7}$$

and $\boldsymbol{w}_{\text{ref}}$ a reference weight vector chosen by the operator and constant in time. The parameter λ affects the distance between $\boldsymbol{w}$ and the reference vector $\boldsymbol{w}_{\text{ref}}$ (usually set

to zero or to $[1/M, \ldots, 1/M]^\top$, following the ensemble mean). The function β_γ gives higher importance to the most recent timesteps. When both λ and γ are set to zero, a simple recursive least-square regression is achieved.

The classical ridge regression without discount provides the theoretical guarantee that the final score of the aggregated forecast will be close to the final score of the best constant linear combination. Indeed we have

$$\sum_{t=1}^{T} \frac{1}{T}(y_t - \widehat{y}_t)^2 - \min_{\boldsymbol{u}\in\mathcal{B}_M} \sum_{t=1}^{T} \frac{1}{T}\left(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right)^2 \leqslant \mathcal{O}\left(\frac{\ln T}{T}\right), \tag{2.8}$$

under the assumption of bounded losses $(y_t - \widehat{y}_t)^2$. The condition $\boldsymbol{u} \in \mathcal{B}_M$, where $\mathcal{B}_M$ is a 2-norm ball in $\mathbb{R}^M$, means that $\|\boldsymbol{u}\|_2^2$ is bounded. The best linear combination with constant weights (implicitly defined above by the second term in the left hand side) is named the oracle. The oracle is found by least-square regression over the whole set. By definition, the oracle shows better performance than any member in the ensemble. Consequently the aggregated forecast is more skillful in the long run than the best member. The discounted ridge regression provides asymptotically this guarantee, which is verified for each sequence of discounted regret.

2.3.3 Algorithm

Parameters: λ, γ, $\boldsymbol{w}_{\text{ref}}$;

Initialization: $\boldsymbol{w}_1$;

For each time index t = 1, 2, ..., T

1. get the predictions $\boldsymbol{x}_t$,
2. compute $\widehat{y}_t$, with $\boldsymbol{x}_t$ and $\boldsymbol{w}_t$,
3. get the observation y_t and compute $\boldsymbol{w}_{t+1}$.

The initial weight vector $\boldsymbol{w}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$.

2.4 Application

2.4.1 Experiment setup

The sequential aggregation with discounted ridge regression as described in 2.3.2 is applied independently at each location of the 127×187 grid. In a similar fashion to the study of Section 2.2, the forecast variable is the incident radiation flux integrated between 06:00 and 12:00 UTC. The TIGGE forecasts are available at 18:00 or 24:00 on day D-1 to forecast the quantity of interest for day D at 12:00, also named (D, 12:00).

The ensemble data are SSR data from TIGGE without any SSR-SSRD conversion because the aggregation does not depend on any multiplicative coefficient applied to the members. In other words, the weights of the aggregation produce multiplicative corrections and solve the issue of the nature of TIGGE data.

The aggregation parameters (λ, γ) are respectively set to 6×10^6 and 20 by default (see Section 2.4.2 for the assessment of the parameters). The values of the reference

vector $\boldsymbol{w}_{\text{ref}}$ are set to $1/M$ in order to drive the aggregated forecast towards the ensemble mean. Even though the ensemble mean may not be the most appropriate reference vector, the vector $\boldsymbol{w}_{\text{ref}}$ mostly impacts the beginning of the aggregation so that the critical parameters are truly (λ, γ).

The aggregation may be achieved in a single step by choosing members from all of the center ensembles. Another approach involves two steps: a first aggregation within each center ensemble and a second aggregation with the resulting forecasts as members. Both procedures have been tested and only the procedure achieved in a single step is presented here since the second procedure did not lead to significantly different results.

In order to study the impact of the number of members M, the same amount of members are chosen from each meteorological center, but the center ensembles do not have the same size. Therefore, the members are chosen in a way that their ranks are linearly spaced and centered on the median of each center ensemble. The full sorting procedure is not impacted by this member selection because in this case the members are mixed up before sorting and rank selection. It is possible to realize aggregation with one single member, such as the reference forecast. In this case the weight plays the role of a local correction factor.

Missing data (only CPTEC ensemble) are replaced by the ensemble mean of the available members.

2.4.2 Results

Aggregation example with one center ensemble

In this section, the aggregation is first run with the ensemble KMA only (without sorting) and then with an additional deterministic member. In the first case, we indicate that the RMSE of the SSRD ensemble mean is $75.0\,\mathrm{W\,m^{-2}}$, whereas the RMSE of the ensemble mean with linear conversion is $72.1\,\mathrm{W\,m^{-2}}$. The RMSE of the aggregated forecast equals $70.0\,\mathrm{W\,m^{-2}}$ and may decrease when data from another source is included in the ensemble. For example, when the ECMWF deterministic forecast is added as a member, the RMSE reaches $67.5\,\mathrm{W\,m^{-2}}$. The scores are to be compared to the $58.9\,\mathrm{W\,m^{-2}}$ RMSE of the oracle of the same ensemble and to the $65.1\,\mathrm{W\,m^{-2}}$ RMSE of the oracle with only KMA data.

One question is whether the improvement due to the deterministic member originates from its high spatial resolution. We therefore include the ECMWF deterministic forecast at the lower resolution of $0.25° \times 0.25°$ which is obtained by averaging the fine $0.125° \times 0.125°$ forecast. In that case, the RMSE is also equal to $67.5\,\mathrm{W\,m^{-2}}$, which means that the resolution of the reference forecast is not the key factor. For comparison, the RMSE of the reference forecast corrected by discounted ridge regression is equal to $68.0\,\mathrm{W\,m^{-2}}$.

Oracles with orthogonal members

We now want to quantify the potential improvements brought by sequential aggregation with all members, using a relevant oracle. The problem of overfitting can arise because of the large number of members (158) compared to the length $T = 350$ of the

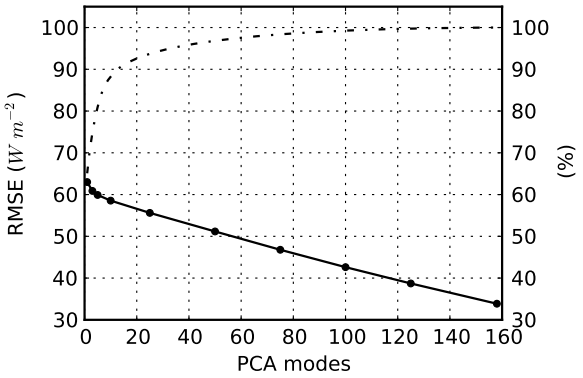


Figure 2.6 – RMSE (solid line with disks, left axis) and explained variance (dashed-dotted line, right axis) of oracles with orthogonal members plotted against number of orthogonal members.

time sequence. In that case, the score of the oracle is artificially good, and the competition against the oracle does not carry any meaning. We present here oracles computed with the help of principal component analysis (PCA), so as to avoid overfitted oracles. The PCA generates orthogonal modes, and consequently orthogonal members. The 158 members are sorted within each center ensemble and then orthogonalized by PCA, independently at each grid point. We compute the oracle of the ensemble of size M', by selecting M' orthogonal members based on the M' modes explaining the largest variance share of the ensemble. In Figure 2.6 the RMSE of the oracle and the total amount of variance explained by its members are plotted against the number of orthogonal members. An indication of possible overfitting is shown in Figure 2.6, because the RMSE still decreases with the number of PCA modes while the share of unexplained variance is small. Indeed the 32 first orthogonal members explain 95% of the variance and generate an oracle with an RMSE of $54.4\,\mathrm{W\,m^{-2}}$, whereas the RMSE of the oracle with all PCA modes (100% of explained variance) equals $33.9\,\mathrm{W\,m^{-2}}$. Consequently, the 32-member oracle is considered as the relevant oracle; its score is a relevant evaluation of the best score that may be achieved by linear combination without overfitting.

Typical aggregation

In this paper, the "typical aggregation" refers to the aggregation with 30 members (5 per sorted center ensemble), with default parameters $\lambda = 6 \times 10^6$ and $\gamma = 20$. The resulting forecast is simply referred to as typical aggregated forecast and is used below to illustrate spatial and temporal features resulting from sequential aggregation. Note that the typical aggregation is compared to all members in Figure 2.13 and to the ensemble means for each month in Figure 2.3. We highlight the fact that the typical aggregated forecast performs better than any of the individual members, steadily over time.

The maps of scores of RMSE, MAE and relative RMSE are shown in Figure 2.7. The MAE and RMSE maps show very similar patterns. The relative MAE is not shown for it is similar to the relative RMSE. The maps indicate that the topographic relief of

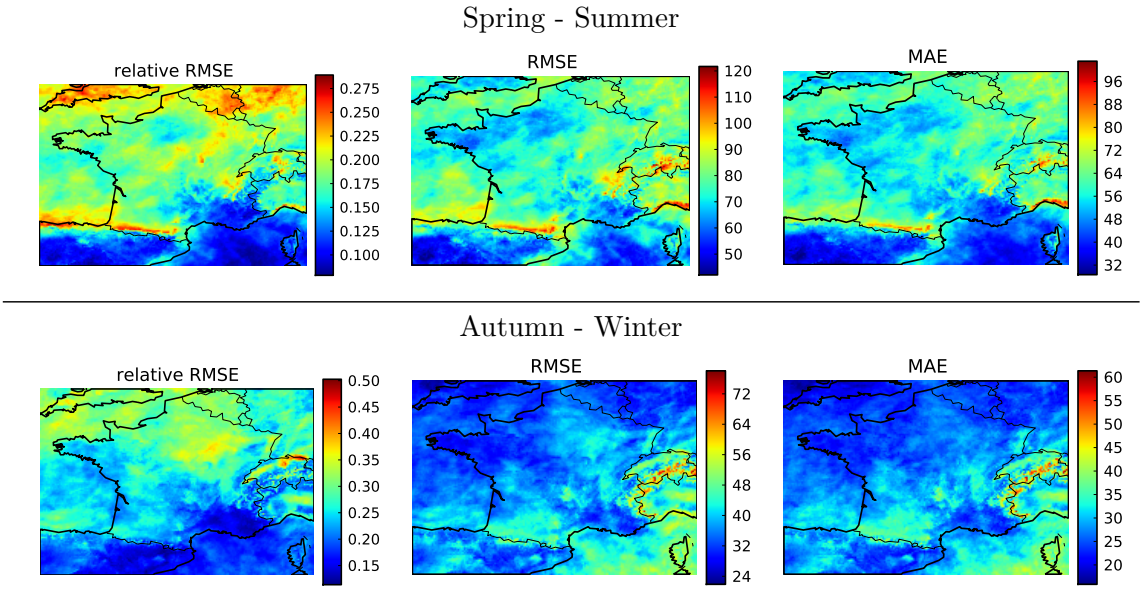
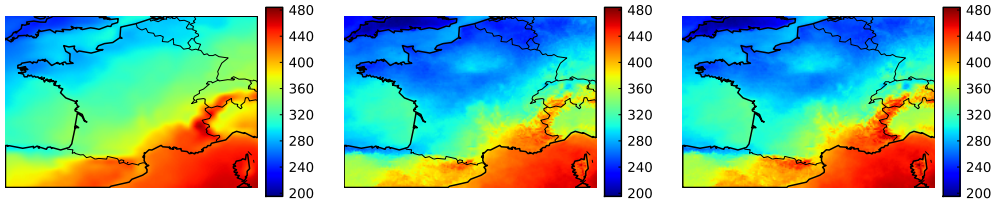


Figure 2.7 – Maps of statistical scores for the typical aggregated forecast (unitless for rRMSE, in $\mathrm{W\,m^{-2}}$ for MAE and RMSE).



(a) ECMWF SSRD deterministic forecast.

(b) Typical aggregated forecast.

(c) HelioClim.

Figure 2.8 – Annual averaged estimation for 12:00 UTC in $\mathrm{W\,m^{-2}}$. Although it is based on individual members with low resolution, the aggregated forecast shows fine structures that are comparable to those of HelioClim.

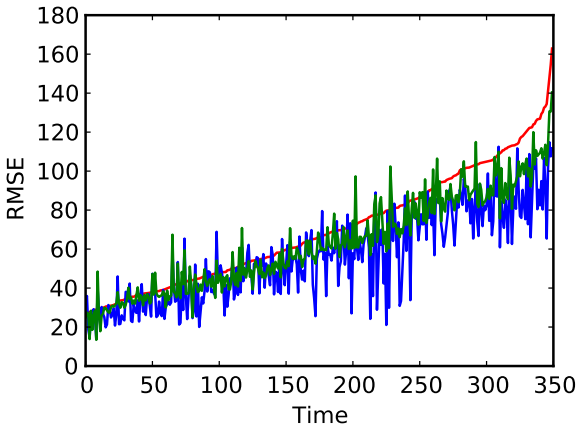


Figure 2.9 – Sorted spatial RMSE in $\mathrm{W\,m^{-2}}$ along 350 days; ECMWF deterministic forecast (red and used to sort the days), typical aggregated forecast (blue), ECMWF deterministic forecast corrected by discounted ridge regression (green).

mountains severely degrades the forecast accuracy. Indeed the absolute largest errors are found in the region of the Alps and in the region of the Pyrenees. The coarse resolution of the members may explain the poor performance in the regions of complex terrain. Also, in the mountainous regions, the albedo shows higher temporal variations, which is difficult to catch for the aggregation. On the opposite, the best scores are steadily achieved in the inner lands of Spain and in the south-east of France. High relative errors are numerous in the northern area especially during the spring-summer period, because of the low values of the observations combined with large errors. For example in spring and summer period, the British area shows a relative RMSE higher than 20%, whereas the south-east of France reaches a relative RMSE below 12.5%, even though the two areas are associated with similar RMSEs over the same period. We recall here that the relative RMSE of the satellite observations compared to BSRN stations is worth 10.7% on average.

Most of the members are computed with a low spatial resolution and do not show fine long-term spatial structures. On the contrary, the aggregated forecast is built independently at each location, which allows the procedure to adapt locally and to finally show fine structures, that are resolved by HelioClim. These structures are finer than any structures found in the ensemble members (due to their low resolutions), even in the ECMWF deterministic forecasts (Figure 2.8).

In order to compare predictions performance at each date, the spatial RMSE of various predictions are temporally sorted according to the performance of the ECMWF reference forecast (Figure 2.9). The overall trend is the same for the three forecasts. The aggregated forecast often shows the best performance, but it does not always perform better than the aggregation applied to the reference forecast only. One benefit of aggregation is that large improvements are achieved at the most difficult timesteps, even with one single member in the ensemble.

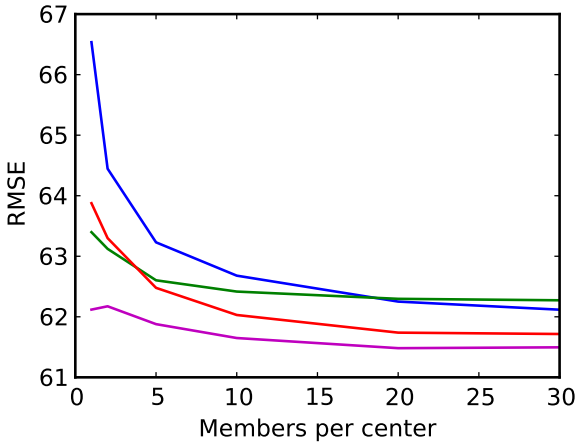


Figure 2.10 – RMSE in $\mathrm{W\,m^{-2}}$ against the maximal number of members chosen from each center ensemble; (blue: no sorting), (green: sorting full ensemble), (red: sorting by center), (magenta: sorting by center and reference forecast included in the ensemble).

Members selection and sorting

The performance of the aggregated forecast is impacted by the number of members up to a certain extent (Figure 2.10). Beyond a total amount of roughly 60 members (10 members from each center ensemble), the scores reach a plateau. The full sorting procedure is the first to reach its own plateau, which may be caused by the merger process of the members with sorting. Besides, the sorting procedures impact the scores with limited consequences (less than $1\,\mathrm{W\,m^{-2}}$) for large enough ensembles. According to the observed performance, the sorting procedure by center should be preferred if more than 5 members per center are chosen. It is noticeable that the reference forecast brings a score improvement whatever the number of aggregated members. The best score achieved with default (λ, γ) equals $61.5\,\mathrm{W\,m^{-2}}$ for the RMSE and $43.6\,\mathrm{W\,m^{-2}}$ for the MAE. Compared to the scores of reference forecast, the improvements brought by the aggregation equal 21.2% for the RMSE and 19.8% for the MAE. Besides, compared to the scores of TIGGE-mean, these improvements are worth 8.1% for the RMSE and 9.1% for the MAE.

The aggregation ensemble may also be built with only 5 out of 6 center ensembles. We compare the aggregated forecast of 30 members, so that 6 members are chosen in each of 5 center ensembles. In this case, the reference aggregated forecast is the typical aggregated forecast with 5 members chosen in each of the 6 center ensembles. Either way, the members are sorted per ensemble. We found that the aggregated forecast with 5 centers performs better than the typical aggregated forecast, when the omitted members are from CMA or CPTEC. The largest improvement occurs when the CPTEC is excluded with a benefit of only $0.3\,\mathrm{W\,m^{-2}}$. On the opposite, the score is the most severely degraded when the members from KMA are left out, generating the RMSE of $64.1\,\mathrm{W\,m^{-2}}$, while the RMSE of the typical aggregated forecast equals $62.5\,\mathrm{W\,m^{-2}}$. To conclude, not using all center ensembles to build the aggregation ensemble can generate both benefits and loss in terms of score. However, these benefits are lower than the loss

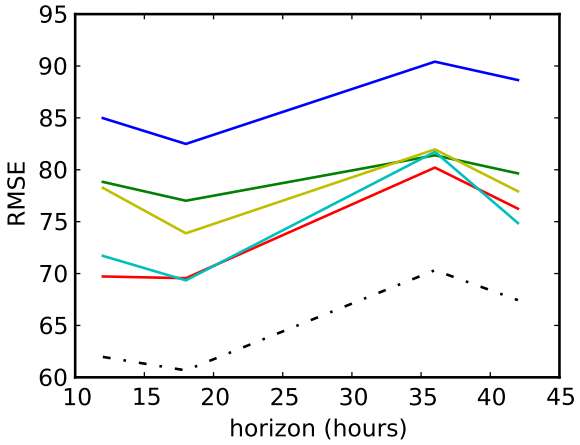


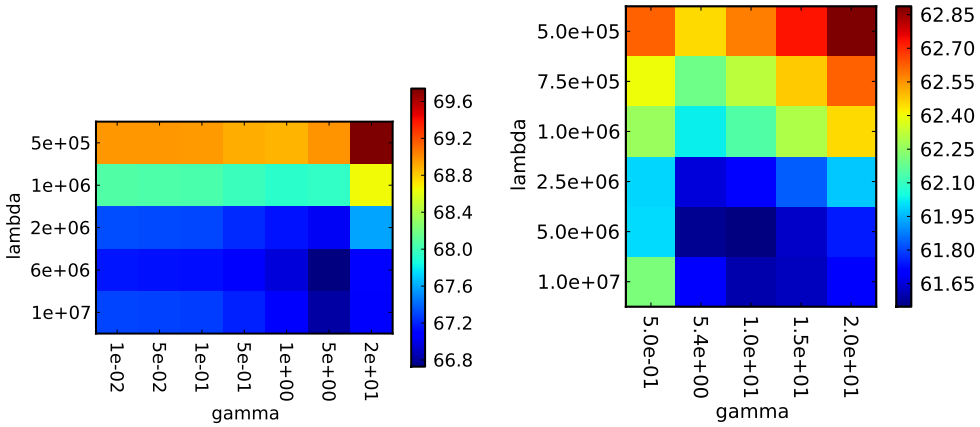
Figure 2.11 – Impact of the horizon up to 42 h on the RMSE of the aggregated forecast and ensemble mean (lin); (dash dot: aggregated with 60 members), (blue: CMA), (green: ECMWF), (red: UKMO), (cyan: KMA), and (yellow: Météo-France).

caused by leaving out skillful forecasts. In practice, one may arguably keep all center ensembles.

Time horizon

We consider an ensemble forecast delivered at time t for the k next steps. The weights of the sequential aggregation are commonly used for the first step ahead (see the algorithm, in Section 2.3.3), but can also be used for the following timesteps $t+k$, where $k > 1$. In fact, the weights can even be used for timesteps that are not included in the sequential aggregation, such as the forecast for 18:00 UTC. This new framework allows us to generate the aggregated forecast for (D, 18:00), (D+1, 12:00) and (D+1, 18:00) with the weights initially computed for the prediction (D, 12:00). In terms of time horizon, the prediction for (D, 12:00) corresponds to the 12 h horizon, (D, 18:00) to 18 h, (D+1, 12:00) to 36 h and (D+1, 18:00) to 42 h. The time horizons of Météo-France forecasts are actually 6 h longer, because its predictions start at 18:00 and not 24:00. The night time predictions steps are skipped because of their very low values.

The RMSE of the ensemble means and the scores the typical aggregated forecast are shown in Figure 2.11, depending on the time horizon. The scores of CPTEC ensemble mean are not shown, because of their high values. The six presented forecasts share the same trends. The scores of predictions for 18:00 beat the scores for predictions 12:00 of the same day, by less than $4\,\mathrm{W\,m^{-2}}$ on average, which is consistent with the fact that the forecasts for 12:00 have higher values than the forecasts for 18:00. Furthermore, the predictions for day D+1 show an average degradation of more than $5\,\mathrm{W\,m^{-2}}$ compared to the prediction for day D. Additionally, we notice that the ensemble mean from ECMWF shows the steadiest performance over time and that the benefits of the aggregated forecast are retained for longer time horizons.



(a) Data set of year 2011 (RMSE for 100 random locations).

(b) Data set of year 2012 (RMSE for all locations).

Figure 2.12 – RMSE grid ($\mathrm{W\,m^{-2}}$) depending on the aggregation parameters, aggregating a 134-member ensemble using at most 30 sorted members from each center ensemble.

Influence of the aggregation parameters

We searched the best set of parameters (λ, γ) that provides the lowest RMSE for the data set of year 2011, already presented in Section 2.2.2. We found that at most $3\,\mathrm{W\,m^{-2}}$ were to be gained in terms of RMSE on a wide range of parameters (Figure 2.12(a)). Our default parameters (6.10^6, 20) produce the score of $67.1\,\mathrm{W\,m^{-2}}$, which is close to the best score achieved on the grid of Figure 2.12(a) (e. g. $66.7\,\mathrm{W\,m^{-2}}$ with $\lambda = 6.10^6$ and $\gamma = 5$). Furthermore, we found that the best set of parameters varies in terms of space and time. Indeed, if we exclude the 50 first timesteps in the RMSE, then the most appropriate parameters are $\lambda = 10^7$ and $\gamma = 20$. If we choose a posteriori the optimal (λ, γ) for each grid point, only $1\,\mathrm{W\,m^{-2}}$ is to be gained from default parameters. The minor variations of performance guarantee that we may choose our default parameters within one order of magnitude, and test them on another data set without significant loss of performance.

We produce the same analysis for the data set of year 2012 in Figure 2.12(b). In this case, the best parameters are $\lambda = 5.10^6$ and $\gamma = 10$, with the score of $61.5\,\mathrm{W\,m^{-2}}$ that is very similar to the score of $61.7\,\mathrm{W\,m^{-2}}$ obtained with default parameters. Therefore, the order of magnitude of the parameters for year 2012 can be deduced from the data set of year 2011. Besides we also found that if we choose the best (λ, γ) for each grid point, the gain does not exceed $1\,\mathrm{W\,m^{-2}}$. As stated with the data set of year 2011, once the relevant order of magnitude for the parameters (λ, γ) is known, only little improvement is possible by adjusting them.

2.5 Conclusion

The ensemble forecasts from TIGGE provide a wide range of meteorological fields including net short-wave solar radiation, with the large timestep of 6 h. After conversion based on a constant albedo, the resulting ensembles are under-dispersed, even when grouped together. The performance of the ensemble means is assessed with RMSE and MAE and compared to the deterministic forecast from ECMWF. The reference forecast has higher resolution than the best TIGGE ensemble means, but produces similar scores. Sequential aggregation brings improvements to the TIGGE ensembles on several features, including RMSE and MAE, with theoretical guarantees. In this study, the aggregated forecast performs better than any member of the ensembles and any ensemble mean. On average, the aggregated forecast retrieves all spatial patterns, even at a much finer resolution than any of the members. Besides, the members combination proves to be consistent with the time horizon. Finally sequential aggregation is easy-to-use for its parameters do not need accurate values.

Practical applications of the aggregation algorithm should investigate higher temporal resolutions, especially the hourly timestep. Next developments may focus on the study of uncertainty with sequential aggregation, which is possible using filtering [MNZ13], but without the same theoretical robustness as in this paper. The introduction of multiple model runs per day with temporal interpolations may result in a robust framework for intraday forecasting. Furthermore, new procedures are to be developed to explore sequential aggregation with spatial dependencies between grid points, with the objective of better forecasting daily spatial patterns. There is also a need for quantifying the possible resolution improvements brought by the aggregation.

Appendix 2.A Conversion from SSR to SSRD and reference forecast

2.A.1 Methods

Several conversion methods allowing the estimation of the albedo are tested. Even though the KMA data set is already SSRD and does not need data conversion, the below methods are also tested on the KMA data set, for the sake of completeness.

First, we infer three conversions from our reference forecast. The SSR and SSRD forecasts from ECMWF are used to provide three SSR-SSRD empirical conversions (Table 2.3): with a constant 1.18 coefficient (glob), with local multiplicative coefficients (mult), and with local additive coefficients (add). The local multiplicative conversion is in fact the local estimation (in space and time) of the factor $1/(1-\alpha)$ based on ECMWF continuous forecast.

Although it may lead to local large errors due to seasonal changes in snow cover and vegetation, we model a constant albedo in space and time (glob). The linear relationship (Eq. (2.1)) between SSR and SSRD forecasts from ECMWF is found by linear regression on data from 100 random locations and for 350 days in year 2012. The resulting slope value (supposed to be equal to $1/(1-\alpha)$) is 1.18 and the intercept value equals $13\,\mathrm{W\,m^{-2}}$. The squared correlation coefficient R^2 of 0.968 shows that

Label	Conversion formula
glob	$x_{tigge} \times 1.18$
mult	$x_{tigge} \times \frac{x^{ssrd}_{ecmwf}}{x^{ssr}_{ecmwf}}$
add	$x_{tigge} - \left(x^{ssr}_{ecmwf} - x^{ssrd}_{ecmwf}\right)$
lin	$x_{tigge} \times \mathrm{a_{center}} + \mathrm{b_{center}}$

Table 2.3 – Empirical conversion formula of TIGGE data x_{tigge} from SSR to SSRD. The deterministic forecasts from ECMWF are named x_{ecmwf}.

Center	Slope $\mathrm{a_{center}}$	Intercept $\mathrm{b_{center}}$ ($\mathrm{W\,m^{-2}}$)	$\mathrm{R^2}$
CMA	1.11	18	0.72
ECMWF	1.18	−21	0.78
UKMO	1.10	19	0.81
KMA	0.91	18	0.81
CPTEC	0.82	41	0.48
Météo-France	1.07	37	0.76

Table 2.4 – Linear regression SSR-SSRD, based on data from year 2011.

SSR and SSRD forecasts are strongly correlated in practice. According to the ratio between ECMWF SSRD and SSR forecasts, the relative standard deviation (standard deviation divided by mean) of the coefficient $1/(1-\alpha)$ is equal to 8.8% on average over all the grid points. Consequently the local variations of the ratio $1/(1-\alpha)$ are rather small compared to its local mean value. Besides, more than 90% of the values of the coefficient $1/(1-\alpha)$ are comprised between 1.13 and 1.48.

Second, a constant albedo is computed for each center ensemble. This method is referred to as "lin" in Table 2.3 and is also referred to as the linear conversion below, because the method is based on linear regressions. Taking HelioClim observations as SSRD data and TIGGE ensemble means as SSR data, linear regressions are carried out on past data sets of year 2011 described in Section 2.2.2. The results of the regressions (Table 2.4) show two slopes lower than one (KMA and CPTEC), and intercepts ranging from $-21\,\mathrm{W\,m^{-2}}$ to $41\,\mathrm{W\,m^{-2}}$. Remind that KMA data is already SSRD data so that its slope is not related to the albedo. The diversity of slopes and intercepts values suggest that the albedo coefficient should be evaluated independently for each center ensemble.

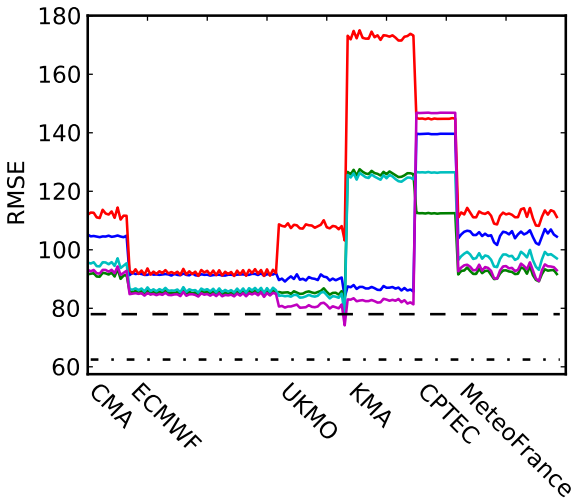


Figure 2.13 – Annual RMSEs of all members in $\mathrm{W\,m^{-2}}$ according to SSR-SSRD conversion method; (blue: no conversion), (green: add), (red: mult), (cyan: glob) and (magenta: lin). The abscissa is the member label and the members are grouped by origin. The dashed line is the score of the reference forecast. The dashed-dotted line is the score of a typical aggregated forecast (see Section 2.4.2).

2.A.2 Numerical results

The deterministic forecasts from ECMWF provide several scores. We build another deterministic forecast, called SSRD-lin, based on deterministic SSR and the linear conversion method. Compared to HelioClim, the RMSE of the ECMWF SSRD deterministic forecast equals $78.0\,\mathrm{W\,m^{-2}}$ while the RMSE of SSRD-lin equals $76.9\,\mathrm{W\,m^{-2}}$. When the largest error percentiles of SSRD and SSRD-lin forecasts are compared, we see that the largest errors belong to SSRD-lin only beyond the 99^{th} percentile. On the one hand, the physical approach with local albedos (as opposed to SSRD-lin) does not generate the largest errors. On the other hand, local albedos do not provide the best score.

Furthermore, the resolution of the ECMWF deterministic forecast does not impact its performance. Indeed the ECMWF deterministic forecast at degraded resolution of $0.25° \times 0.25°$, by interpolation, shows the RMSE of $77.7\,\mathrm{W\,m^{-2}}$ when compared to HelioClim, which is below the RMSE of the reference forecast. The two predictions at different resolutions have similar performance, since the distance between them, measured with a root mean square discrepancy (in time and space), equals $3.0\,\mathrm{W\,m^{-2}}$.

The RMSEs are shown in Figure 2.13 for all members without sorting procedures. The RMSEs between individual members and satellite observations are at least twice as big as the RMSEs between satellite observations and ground measurements (Table 2.2). It is noteworthy that members of a same center ensemble perform similarly. The scores strongly depend on the conversion methods and on the origin of each forecast. We highlight again the fact that the nature of KMA data is already SSRD.

The linear conversion (lin) is the best conversion method tested here except for the

CPTEC ensemble. The multiplicative conversion method was supposed to provide accurate forecasts. However, here is confirmed the need for the conversion method to adapt to each center ensemble.

3 Online learning with the CRPS for ensemble forecasting

Ensemble forecasting resorts to multiple individual forecasts to produce a discrete probability distribution which accurately represents the uncertainties. Before every forecast, a weighted empirical distribution function is derived from the ensemble, so as to minimize the Continuous Ranked Probability Score (CRPS). We apply online learning techniques, which have previously been used for deterministic forecasting, and we adapt them for the minimization of the CRPS. The proposed method theoretically guarantees that the aggregated forecast competes, in terms of CRPS, against the best weighted empirical distribution function with weights constant in time. This is illustrated on synthetic data. Besides, our study improves the knowledge of the CRPS expectation for model mixtures. We generalize results on the bias of the CRPS computed with ensemble forecasts, and propose a new scheme to achieve fair CRPS minimization, with essentially no assumption on the distributions.

A slightly different version of this chapter was published as the paper: Thorey, J., Mallet, V., and Baudin, P. « Online learning with the CRPS for ensemble forecasting ». In: *Quarterly Journal of the Royal Meteorological Society* (2016).

Contents

Introduction

The minimization of the CRPS is a common way to drive probabilistic forecasts [Gne+05; JMA15]. From diagnostic tools to modeling techniques, Gneiting and Katzfuss [GK14] review the state of the art of probabilistic forecasting. Using several forecasts, based on various models and perturbed input data, is a common way to produce probabilistic forecasts [LP08]. The roots of this framework known as ensemble forecasting is reviewed by Lewis [Lew05]. Ensemble of forecasts is the raw material of the techniques proposed in this paper.

Sequential aggregation targets optimal combinations, as thoroughly introduced in the monograph Cesa-Bianchi and Lugosi [CL06]. These techniques, also known under the scope of online learning, come with attractive theoretical guarantees of performance. Stoltz [Sto10] and Mallet et al. [MSM09] and Mallet [Mal10] summarized and tested these techniques on forecasts of respectively electricity consumption and ozone concentrations. Usually focused on scalar forecasting, sequential aggregation was applied to the Brier score and the quantile score by respectively Vovk and Zhdanov [VZ09] and Biau and Patra [BP11]. In this paper, we use sequential aggregation in order to target the best CRPS, with theoretical guarantees that require essentially no assumption on the forecast or observation distributions. In this sense, our method is a non-parametric post-processing method. Our techniques generate weights for each ensemble member so as to produce a linear opinion pool, also known as model mixture [GM90; CW99]. Ranjan and Gneiting [RG10] provide mathematical grounds on these combinations. Our technique was first designed to work with an ensemble of scalar forecasts. Still, it can be applied when a parameterized distribution is associated to each forecast.

In Section 3.1, we describe the mathematical background on the CRPS. We provide contributions related to ensemble forecasting with discrete Cumulative Distribution Functions (CDFs). Our contributions are mainly generalizations of existing results to the case of combinations of forecasts with unequal weights, in a probabilistic framework. We also provide a framework to work with classes of members, compatible with fair probabilistic evaluations. In Section 3.2, we detail online learning techniques, with adaptation for probabilistic ensemble forecasting based on the CRPS. In Section 3.3, we illustrate the notions of Section 3.1 with numerical experiments, and we demonstrate our algorithms with numerical examples. We summarize several useful identities involving CDFs in Appendix 3.A.

3.1 Mathematical background

3.1.1 Bibliographical remarks

The evaluation of probabilistic forecasts is a long range discussion going on since Winkler and Murphy [WM68] and Savage [Sav71]; see Dawid [Daw08] for a detailed bibliographical analysis, and more recently Gneiting and Raftery [GR07] and Candille and Talagrand [CT05] for detailed analyses. The Brier score was introduced by Brier [Bri50] and Good [Goo52] to evaluate probabilistic forecasts for a given threshold and a binary observation. The Continuous Ranked Probability Score (CRPS) can be viewed

as a continuous version of the Brier score [Eps69; Mur71] for any threshold.

3.1.2 The Continuous Ranked Probability Score (CRPS)

We want to forecast a scalar quantity y called the verification and we suppose that y admits an underlying distribution that is described by the CDF F. The CRPS is considered as a realization of a random variable, and it is defined as

$$\mathrm{CRPS}(\mathrm{G}, y) = \int (\mathrm{G} - \mathrm{H}_y)^2 \,, \tag{3.1}$$

where G is a CDF that is chosen by the forecaster in order to predict F, H is the unit (or Heaviside) step function, and $\mathrm{H}_y(x)$ indicates a centered Heaviside function $\mathrm{H}(x - y)$. The CRPS is negatively oriented, meaning that lower scores imply better performance. Gneiting and Raftery [GR07] show that the CRPS may also be written as

$$\mathrm{CRPS}(\mathrm{G}, y) = \mathrm{E}(|\mathrm{X} - y|) - \frac{1}{2}\,\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)\,, \tag{3.2}$$

where E is the expectation, and both X and X' are two random variables drawn from G. A decomposition of the average CRPS was introduced by Hersbach [Her00]. The decomposition of scores into divergence and uncertainty terms is explained in Bröcker [Brö09]. The average CRPS is decomposed as follows:

$$\int \mathrm{CRPS}(\mathrm{G}, y)\mathrm{dF}(y) = \int (\mathrm{G} - \mathrm{F})^2 + \int \mathrm{F}(1 - \mathrm{F})\,, \tag{3.3}$$

where y is integrated over the values defined by F (using Equation 3.28 of Appendix 3.A). The CRPS is a strictly proper score, which means that it is minimized on average if and only if the forecaster's choice G is equal to F. This is a straightforward observation from Equation 3.3.

The strict propriety of the CRPS can be compared to the non-strict propriety of the square loss [BS07b], which reads

$$(y - \mathrm{E}(\mathrm{X}))^2 = \left(\int \mathrm{G} - \mathrm{H}_y\right)^2 \,, \tag{3.4}$$

according to Equation 3.31 of Appendix 3.A. We see that minimizing the square loss and minimizing the CRPS (Equation 3.1) are rather different objectives, due to the location of the square function inside or outside the integral expression. The CRPS objective is more demanding, because in this case the integration is applied to a positive function.

3.1.3 The ensemble CRPS

In the case of ensemble forecasting, the forecaster relies on an ensemble of M members x_m, $m \in \{1, \ldots, M\}$, to construct a CDF. The empirical CDF $\mathrm{G}^{\mathcal{E}}$ of the ensemble is a step function with jumps of heights u_m (called weights) at the members values x_m. Thus we write $\mathrm{G}^{\mathcal{E}}(x) = \sum_{m=1}^{M} u_m \mathrm{H}(x - x_m)$. In order to satisfy the definition of a CDF, the weights u_m should be nonnegative and sum to one, so that they produce a

convex combination. Such weight vectors define the simplex $\mathcal{P}_M$ of $\mathbb{R}^M$. The weights u_m are to be optimized in order to minimize the CRPS.

The computation of the integral of Equation 3.1 is easy on step functions $\mathrm{G}^{\mathcal{E}}$. When the CDF is a step function, we refer to the score as the ensemble CRPS:

$$\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) = \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2} \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| \,. \tag{3.5}$$

The derivation of Equation 3.1 is detailed in Appendix 3.B.

Without further information, the members are assumed to be i.i.d., thus the forecaster may arguably choose all weights equal to $1/M$. By definition, a scoring rule depending on the verification y and i.i.d. members x_m is fair if the average score is minimized when the members and the verification are sampled from the same distribution. Ferro et al. [FRW08] show that the ensemble CRPS is unfair due to the finite size of the ensemble. In the next section, we generalize this result to the case of unequal weights, with non identically distributed members.

The bias of the score is an important topic in our optimization framework. Indeed if our objective function is intrinsically biased, then the resulting probabilistic forecast cannot be calibrated.

3.1.4 Bias of the ensemble CRPS with underlying mixture model

We consider that the members x_m are independent samples from the CDFs G_m, and that y is fixed. The purpose of this section is to compare the score obtained with the step function $\mathrm{G}^{\mathcal{E}}$ averaged according to the CDFs G_m and the score obtained with the mixture model described by the average CDF $\mathrm{G} = \sum u_m \mathrm{G}_m$.

Taking the expectation with respect to the members x_m, we show that

$$\mathrm{E}(\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y)) = \int \mathrm{H}_y - 2 \sum_{m=1}^{M} u_m \mathrm{G}_m \mathrm{H}_y + \sum_{m \neq k}^{M} u_m u_k \mathrm{G}_m \mathrm{G}_k + \sum_{m=1}^{M} u_m^2 \mathrm{G}_m \,, \tag{3.6}$$

using Equation 3.28. The trick is that $\mathrm{H}^2(x - x_m) = \mathrm{H}(x - x_m)$, thus the average CRPS does not include G_m^2 terms but G_m terms instead. We conclude by introducing the terms

$\sum_{m=1}^{M} u_m \mathrm{G}_m$ and $\sum_{m=1}^{M} u_m^2 \mathrm{G}_m^2$ in conjunction with Equation 3.29, 3.33 and 3.34:

$$\begin{aligned}
\mathrm{E}(\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y)) &= \int \mathrm{H}_y + \sum_{m=1}^{M} u_m \mathrm{G}_m - 2 \sum_{m=1}^{M} u_m \mathrm{G}_m \mathrm{H}_y \\
&\quad + \sum_{m \neq k}^{M} u_m u_k \mathrm{G}_m \mathrm{G}_k - \sum_{m=1}^{M} u_m \mathrm{G}_m + \sum_{m=1}^{M} u_m^2 \mathrm{G}_m \\
&= \mathrm{E}(|\mathrm{X} - y|) + \int \sum_{m \neq k}^{M} u_m u_k \mathrm{G}_m \mathrm{G}_k - \sum_{m=1}^{M} u_m \mathrm{G}_m + \sum_{m=1}^{M} u_m^2 \mathrm{G}_m \\
&= \mathrm{E}(|\mathrm{X} - y|) + \int \sum_{m,k}^{M} u_m u_k \mathrm{G}_m \mathrm{G}_k - \sum_{m=1}^{M} u_m \mathrm{G}_m + \sum_{m=1}^{M} u_m^2 \mathrm{G}_m - \sum_{m=1}^{M} u_m^2 \mathrm{G}_m^2 \\
&= \mathrm{E}(|\mathrm{X} - y|) - \frac{1}{2} \mathrm{E}(|\mathrm{X} - \mathrm{X}'|) + \frac{1}{2} \sum_{m=1}^{M} u_m^2 \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|) \,.
\end{aligned}$$

where X and X_m are random variables with CDFs G and G_m respectively. Therefore, we have

$$\mathrm{E}(\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y)) = \mathrm{CRPS}(\mathrm{G}, y) + \frac{1}{2} \sum_{m=1}^{M} u_m^2 \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|) \,. \tag{3.7}$$

In the expectation of the ensemble CRPS, the diagonal terms $u_m^2 \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|)$ are missing, because the spread of each member is assumed to be null. The absence of the diagonal terms is the cause of the bias of the ensemble CRPS. One consequence of Equation 3.7 is that the expected ensemble score is never smaller than the score obtained with G. This fact also follows from the convexity of the CRPS, as pointed out by an anonymous reader.

As a consequence, the minimization of the ensemble CRPS should not be targeted because the solution of this optimization problem is not the underlying CDF of the verification. There is no contradiction with the strict propriety of the CRPS because, for the ensemble CRPS, the solution is only searched in a subspace made of step functions.

In the case of equal weights with i.i.d. members, Fricker et al. [FFS13] detailed in their Appendix why minimizing the ensemble CRPS is misleading as stated above. Ferro et al. [FRW08] exhibit a fair adjusted CRPS score, which includes correction terms to counteract the bias:

$$\mathrm{CRPS}_a(\mathrm{G}^{\mathcal{E}}, y) = \frac{1}{M} \sum_{m=1}^{M} |x_m - y| - \frac{1}{2} \sum_{m,k=1}^{M} \frac{|x_m - x_k|}{M(M-1)} \tag{3.8}$$

$$= \mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) - \frac{1}{2M} \sum_{m,k=1}^{M} \frac{|x_m - x_k|}{M(M-1)} \,. \tag{3.9}$$

We see that rather than being a new score, the adjusted ensemble CRPS is a better estimation of the original CRPS, where the underlying distributions of the members are taken into account. In Equation 3.8, the dispersion of the ensemble $\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)$ is estimated by $\sum_{m,k=1}^{M} |x_m - x_k|/(M(M-1))$. In other terms, the bias terms $u_m^2 \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|)$ of Equation 3.7 are taking into account in Equation 3.9 as $\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)/M^2$, by considering that $\mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|) = \mathrm{E}(|\mathrm{X} - \mathrm{X}'|)$.

3.1.5 Mixture model described by classes of members

We propose in this section a framework compatible with both ensemble forecasting and unbiased scores. In a standard model mixture design, a forecaster will assign weights to known parametric distributions [Raf+05; Gri+06]. We do not want to make assumptions on distributions, thus we use a standard ensemble forecasting framework, where the members are usually assumed to be sampled from unknown CDFs. The goal of this section is to show that despite the finite size of the ensemble, it is possible to use the CRPS by counteracting the discretization-induced bias. This framework is close to what is introduced in Fraley et al. [FRG10], however this previous work focused on Bayesian Model Averaging (BMA), and did not include considerations on the CRPS.

We assume that ensemble members are grouped into classes within which members are i.i.d. In this new setting, a class C has a weight $\mathcal{W}_C$ uniformly distributed among its members. The weight $u_m = \mathcal{W}_C/M_C$ is assigned to the mth member of the ensemble, assuming that it belongs to class C and that class C has M_C members. As an example, classes may be defined according to the rank of the members. Assuming that 10 members are available, two classes may be built by assigning the 5 members with the lowest values to the first class and the remaining members to the second class.

We introduce the CRPS using the classes. We call this score the class CRPS, and denote it

$$\mathrm{CRPS}_{\mathfrak{C}}(\mathrm{G}^{\mathfrak{C}}, y) = \sum_{C\in\mathfrak{C}} \mathcal{W}_C \widehat{\mathrm{E}}(|\mathrm{X}_C - y|) - \frac{1}{2}\sum_{C,D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|)\,. \tag{3.10}$$

The terms of the class CRPS are detailed below.

For the class C, with M_C members x_c^C associated to the random variables X_C and X'_C, we have

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - y|) = \sum_{c=1}^{M_C} |x_c^C - y|/M_C\,, \tag{3.11}$$

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) = \sum_{c=1}^{M_C}\sum_{d=1}^{M_D} |x_c^C - x_d^D|/(M_C M_D)\,, \tag{3.12}$$

where class D is different from class C, and

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) = \sum_{c,c'=1}^{M_C} |x_c^C - x_{c'}^C|/(M_C(M_C-1))\,. \tag{3.13}$$

This last quantity can be seen as the dispersion associated to the i.i.d. members of class C. Note the bias correction of $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ with the factor $M_C(M_C-1)$.

Now we show how the ensemble CRPS and the class CRPS are related. Summing among classes (which belong to the partition $\mathfrak{C}$ of the set of the members), we have

$$\sum_{C\in\mathfrak{C}} \mathcal{W}_C \sum_{c=1}^{M_C} |x_c^C - y|/M_C = \sum_{m=1}^{M} u_m |x_m - y|\,. \tag{3.14}$$

Then we sum inter- and intra-class dispersions to link them to inter-member differences $|x_m - x_k|$. The key point is that inter-member differences for i.i.d. members comprise the intra-class dispersions. We note that

$$\mathcal{W}_C^2 \sum_{c,c'=1}^{M_C} \frac{|x_c^C - x_{c'}^C|}{M_C(M_C-1)} = \frac{M_C}{M_C-1} \sum_{c,c'=1}^{M_C} \left(\frac{\mathcal{W}_C}{M_C}\right)^2 |x_c^C - x_{c'}^C| \,, \tag{3.15}$$

and $M_C/(M_C-1) = 1 + 1/(M_C-1)$ to obtain

$$\begin{aligned}
\sum_{C,D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) &= \sum_{C\neq D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) + \sum_{C\in\mathfrak{C}} \mathcal{W}_C^2 \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{C\in\mathfrak{C}} \frac{1}{M_C-1} \sum_{c,c'=1}^{M_C} \frac{\mathcal{W}_C^2}{M_C^2} |x_c^C - x_{c'}^C| \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{C\in\mathfrak{C}} \frac{\mathcal{W}_C^2}{M_C} \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{m=1}^{M} u_m^2 \widehat{\mathrm{E}}(|\mathrm{X}_{C_m} - \mathrm{X}'_{C_m}|) \,,
\end{aligned}$$

where C_m is the class in which x_m falls. To obtain the last equation, consider that $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ is counted M_C times.

Compared to the ensemble CRPS, the class CRPS admits M additional terms corresponding to the dispersion of each member and resulting from the classes definition:

$$\mathrm{CRPS}_{\mathfrak{C}}(\mathrm{G}^{\mathfrak{C}}, y) = \mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) - \frac{1}{2} \sum_{m=1}^{M} u_m^2 \, \mathrm{E}(|\mathrm{X}_{C_m} - \mathrm{X}'_{C_m}|) \,. \tag{3.16}$$

In the case of a single class, the class CRPS is equal to the adjusted ensemble CRPS described in Section 3.1.4.

The i.i.d. assumption on the members can be seen as too strong. The exchangeability of the members is a relaxation of the i.i.d. assumption. By definition, the joint distribution function of exchangeable members is invariant under permutation of the arguments, thus the members are indistinguishable. We refer the reader to Ferro [Fer14] for an analysis of fair scoring rules with the exchangeability assumption. In a few words, the user must investigate the (generally unknown) dependence structure and taylor the appropriate scoring rule accordingly. The simple case of pairwise uncorrelated members is however tractable. For the ensemble CRPS, the case of pairwise uncorrelated members is in practice equivalent to the case of i.i.d. members, because the terms $|x_m - x_k|$ rely on pairwise correlations only. In the same way for the class CRPS, the assumption of pairwise uncorrelated members within each class and independent members between classes leads to similar results than i.i.d. members. Under the more general assumption of exchangeable members within each class, the definition of $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ should take into account the dependence between members.

Also note that these assumptions are only needed to counter the bias in the ensemble CRPS. Our aggregation methods still remain applicable without such correction. The

theoretical bounds described in the next section do not rely on any stochastic assumption on the prediction data and the verifications. The assumptions of i.i.d. members and the use of the class CRPS should only guide the choice of a loss function.

3.2 Online learning methods

3.2.1 Theoretical background

Up to this section, a single time t was considered. Now we introduce online learning techniques. In this setting, the forecaster receives prediction data $\boldsymbol{\mathcal{D}}_t$ and wishes to produce the best prediction of y_t. In our case, prediction data are ensemble members and the algorithm gives a rule to compute the weights $u_{m,t}$ before each forecast time t. This rule takes into account only past information, and is therefore called the update rule. The goal of a given online learning algorithm is to provide the best possible weights according to a chosen loss function, for example the ensemble CRPS

$$\ell_t^{CRPS_\mathcal{E}}(\boldsymbol{u}) = \int \left(\sum_{m=1}^{M} u_m \mathrm{H}_{m,t} - \mathrm{H}_{y_t} \right)^2 , \tag{3.17}$$

written above for time t. The notation $\ell_t(\boldsymbol{u})$ emphasizes the importance of the weights, as opposed to the ensemble members and the verifications which are assumed to be given to the forecaster.

In practice, the algorithm reads

Initialization: $\boldsymbol{u}_1$;

For each time index $t = 1, 2, ..., T$

1. get prediction data $\boldsymbol{\mathcal{D}}_t$,
2. compute the forecaster's choice with $\boldsymbol{\mathcal{D}}_t$ and $\boldsymbol{u}_t$,
3. get the verification y_t and compute $\boldsymbol{u}_{t+1}$, based on the update rule.

The initial weight vector $\boldsymbol{u}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$.

The performance of an update rule comes with theoretical guarantee, where the forecaster's results are assessed against a reference, which is usually the best forecast with weights constant in time, called the oracle. An important aspect of these theoretical guarantees is that they come with essentially no stochastic assumption on the prediction data and the verifications. In this paper, the theoretical guarantees are regret bounds of the form

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u} \in \mathcal{P}_M} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leqslant o(T) , \tag{3.18}$$

where ℓ_t is assumed to be bounded. The bound of ℓ_t can be arbitrarily small or large, so that this restriction is compatible with essentially all real world applications. Averaging the losses in time (i.e., dividing by T) shows that an algorithm giving the weights $\boldsymbol{u}_t$ is guaranteed to perform at least as well as any mixture model with weights constant in time and based on the same prediction data. This includes any individual forecast and any subset ensemble with uniform weights.

We now consider two algorithms: the online ridge regression and the exponentiated gradient method (EG). We introduce these methods in a general framework, and we show how the methods can be applied to the case of the CRPS. For the algorithm run with ensemble CRPS, a weight is explicitly given to each member. The quantities $|x_{m,t} - y_t|$ and $|x_{m,t} - x_{k,t}|$ are explicitly used in the minimization process. For the algorithm run with class CRPS, equal weights are given to all the members within a class. The weights $\mathcal{W}_{C,t}$ are computed using the terms $\widehat{\mathrm{E}}(|\mathrm{X}_{C,t} - y_t|)$ and $\widehat{\mathrm{E}}(|\mathrm{X}_{C,t} - \mathrm{X}_{D,t}|)$. Combining parameterized distributions is also possible with online learning techniques. It necessitates to compute the quantities $\mathrm{E}(|\mathrm{X}_{m,t} - y_t|)$ and $\mathrm{E}(|\mathrm{X}_{m,t} - \mathrm{X}_{k,t}|)$. These quantities are tractable from the CDFs using Equation 3.29. They are computed in Grimit et al. [Gri+06] for a Gaussian mixture distribution.

3.2.2 Ridge regression

The approach of the ridge regression can be directly expressed in terms of minimization. The update rule for time $t+1$ and based on the loss ℓ is

$$\boldsymbol{u}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{R}^M}{\operatorname{argmin}} \lambda \boldsymbol{w}^\top \boldsymbol{w} + \sum_{t'=1}^{t} \ell_{t'}(\boldsymbol{w}) \,. \tag{3.19}$$

The regularization term with parameter $\lambda \geqslant 0$ controls the 2-norm of the weight vector. It is possible to add discount factors in the sum of the past losses, in order to give more importance to recent timesteps. At first sight, the ridge regression does not constrain the weights to be positive or sum to one. In practice, for the CRPS, we observed that these constraints are approximately satisfied after a spin up period. Other regularization terms of the form $\lambda(\boldsymbol{w} - \boldsymbol{u}_1)^\top(\boldsymbol{w} - \boldsymbol{u}_1)$ may also be used with arbitrary reference vector $\boldsymbol{u}_1 \in \mathcal{P}_M$. The reader interested in recent advances in online regularized regression is addressed to Orabona et al. [OCC15].

For a given experiment length T, for any vector $\boldsymbol{u} \in \mathcal{P}_M$, and if the CRPS losses $\ell_t(\boldsymbol{u}_t)$ are bounded, we have:

$$\mathcal{R}_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leqslant \mathcal{O}\left(\ln T\right) \,, \tag{3.20}$$

so that the so-called regret $\mathcal{R}_T(\boldsymbol{u})$ is sublinear.

The Appendix 3.C details technical aspects, such as the proof for the bound 3.20, as well as guidelines to compute the weights. The ridge regression applied to the square loss $(\mathrm{E}(\mathrm{X}) - y)^2$ gives a similar regret bound in terms of square losses. Our proof for the CRPS is inspired from the proof of the regret bound for the square loss, concisely described by Cesa-Bianchi and Lugosi [CL06]. We were helped by the work of Mallet et al. [MMS07], who demonstrated the case of multiple verification locations (also called stations) for the square loss. Our work transposes the results for the square loss with multiple locations to multiple Brier score with different thresholds, and to the CRPS.

Method	Gradient loss
Ensemble CRPS	$\widetilde{\ell}_{m,t} = \lvert x_{m,t} - y_t \rvert - \sum_{k=1}^{M} u_{k,t} \lvert x_{m,t} - x_{k,t} \rvert + y_t - \sum_{k=1}^{M} u_{k,t} x_{k,t}$
Class CRPS	$\widetilde{\ell}_{C,t} = \widehat{\mathrm{E}}(\lvert \mathrm{X}_{C,t} - y_t \rvert) - \sum_{D \in \mathfrak{C}} \mathcal{W}_{D,t} \widehat{\mathrm{E}}(\lvert \mathrm{X}_{C,t} - \mathrm{X}_{D,t} \rvert) + y_t - \widehat{\mathrm{E}}(\mathrm{X}_t)$
CRPS for general mixture models	$\widetilde{\ell}_{m,t} = \mathrm{E}(\lvert \mathrm{X}_{m,t} - y_t \rvert) - \sum_{k=1}^{M} u_{k,t} \, \mathrm{E}(\lvert \mathrm{X}_{m,t} - \mathrm{X}_{k,t} \rvert) + y_t - \mathrm{E}(\mathrm{X}_t)$

Table 3.1 – Formulae of the loss gradients. Equations from Appendix 3.A are used for the simplifications. The terms of the form $y_t - \mathrm{E}(\mathrm{X}_t)$ do not impact the computation of the weights for EG, because they are independent of the member m or the class C.

3.2.3 Exponentiated gradient

Let the learning rate η be strictly positive, EG follows a multiplicative update rule of the form:

$$u_{m,t+1} = \frac{u_{m,t} \exp\left(-\eta \widetilde{\ell}_{m,t}\right)}{\sum_{m'=1}^{M} u_{m',t} \exp\left(-\eta \widetilde{\ell}_{m',t}\right)}, \tag{3.21}$$

where

$$\widetilde{\ell}_{m,t} = \frac{\partial \ell_t}{\partial u_m}(\boldsymbol{u}_t). \tag{3.22}$$

This update relates to Bayesian inference [Cat04; Aud09]. The algorithm EG admits a formulation in terms of cost function minimization, where the regularization function is the entropy function, also known as the Kullback-Leibler divergence [KW97]. The EG algorithm reads:

$$\boldsymbol{u}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathcal{P}_M} \sum_{m=1}^{M} w_m \ln\left(\frac{w_m}{u_{m,t}}\right) + \eta w_m \widetilde{\ell}_{m,t}. \tag{3.23}$$

Examples of loss gradients are provided in Table 3.1. The loss gradient $\widetilde{\ell}_{m,t}$ of the CRPS has two main terms: (i) $\mathrm{E}(\lvert \mathrm{X}_{m,t} - y_t \rvert)$ accounting for the distance between the verification and the mth random variable $\mathrm{X}_{m,t}$, and (ii) the weighted sum of $\mathrm{E}(\lvert \mathrm{X}_{m,t} - \mathrm{X}_{k,t} \rvert)$ accounting for distances between $\mathrm{X}_{m,t}$ and the $\mathrm{X}_{k,t}$. The first term controls a deviation from the median of the underlying distribution of the verifications, and the second term controls the dispersion of the mixture model. On average (on the verifications), the loss gradients are null if the verifications are correctly described by the forecaster's CDF.

The advantage of using the loss gradients is described (at least) in Devaine et al. [Dev+13]. In a few words, using the loss gradients makes the algorithm competitive against the best convex combination with constant weights, whereas simply using the loss $\ell_{m,t} = \mathrm{E}(\lvert \mathrm{X}_{m,t} - y_t \rvert) - 0.5 \, \mathrm{E}(\lvert \mathrm{X}_{m,t} - \mathrm{X}'_{m,t} \rvert)$ would make the algorithm compete only against the best member. We insist on the fact that using the loss gradients provides the terms $\mathrm{E}(\lvert \mathrm{X}_{m,t} - \mathrm{X}_{k,t} \rvert)$ which are critical for the control of the ensemble spread.

Table 3.2 – Parameters of the numerical experiment.

s_1	s_2	A	B	ω_1	ω_2	T
0.3	0.3	1.68	0.336	1/365.25	1/11	730

The theoretical guarantee for EG states that, if the loss function ℓ is convex with respect to $\boldsymbol{u}$ and admits a subgradient, and if the losses $\widetilde{\ell}_{m,t}$ are bounded within a constant interval $[-a, a]$, then we have:

$$\sup \left[\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u} \in \mathcal{P}_M} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \right] \leqslant \frac{\ln M}{\eta} + \eta \frac{a^2}{2} T \,, \tag{3.24}$$

where the supremum is taken for all possible values of the members $x_{m,t}$ and the verifications y_t, and η is the learning rate [Dev+13]. For optimized values of η proportional to $1/\sqrt{T}$, the regret is sublinear. The theoretical guarantee of Equation 3.24 is verified for the square loss and for the CRPS.

3.3 Numerical example

3.3.1 Simple model

We use the simple model described in Bröcker [Brö12]. The model is supposed to mimic local temperatures. We chose this model because the uncertainty terms are known, consequently we can easily draw conclusions from numerical tests.

We built the verifications y_t from the exact time series

$$a_t = \left(A sin(\pi \omega_1 t) + B sin(\pi \omega_2 t) \right)^2 , \tag{3.25}$$

combined with multiplicative and additive perturbation terms:

$$y_t \sim a_t (1 + s_1 \mathcal{N}(0,1)) + s_2 \mathcal{N}(0,1) \,. \tag{3.26}$$

Each term $\mathcal{N}(0,1)$ represents an independent Gaussian noise with zero mean and a variance of one. The perturbation terms are sampled independently at each timestep. The parameters are summarized in Table 3.2.

The members are sampled as

$$x_{m,t} \sim a_t (1 + s_1 \mathcal{N}(0, d_{ens})) + s_2 \mathcal{N}(0, d_{ens}) \tag{3.27}$$

analogously to the verification distribution, but the standard deviation d_{ens} describing the perturbations terms may differ from its optimal value (i.e., 1). The parameter d_{ens} is also referred to as the dispersion parameter.

3.3.2 Experiments without online learning

In this first experiment, ensembles are built for different values of the dispersion parameter d_{ens}. The members are drawn independently, and the weights of the members

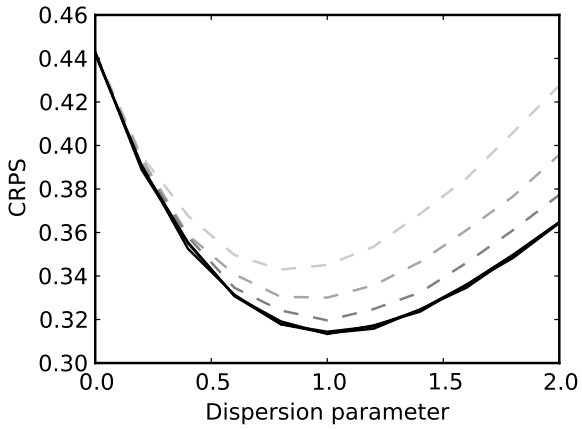


Figure 3.1 – Ensemble CRPS (dotted gray) and adjusted ensemble CRPS (solid black), for ensembles of various sizes (10, 20, 50) from light gray to dark gray. The dispersion parameter (x-axis) is d_{ens}. The scores are averaged over nearly 200 years of data (73,000 timesteps). The (solid black) lines of the adjusted CRPS are approximately at the same location for all ensemble sizes.

are taken constant and all equal to $1/M$. As expected, the adjusted ensemble CRPS gets the lowest value when the ensemble shows the correct spread (i.e., for $d_{ens} = 1$), see Figure 3.1. On the contrary, the best (non adjusted) ensemble CRPS is obtained for under-dispersed ensembles $d_{ens} < 1$. The shift of the ensemble CRPS minimum from the ideal location $d_{ens} = 1$ is larger for ensembles of small size, because the bias of the ensemble CRPS is proportional to $1/M$. This is a direct illustration of the bias due to the limited size of the ensemble explained in Section 3.1.4.

3.3.3 Experiments with weight updates

Now we test online learning techniques and more specifically their ability to discriminate between members. We build an ensemble of $M = 10$ members, that is composed of two subensembles, or classes, of equal size. The first subensemble is defined by the same distribution than the verifications. The second subensemble follows a distribution controlled by d_{ens}. If $d_{ens} = 1$, then the whole ensemble is correctly dispersed. In other words, half of the members follow the correct distribution, while the second half can follow a different distribution.

An example of the temporal evolution of the weights is given in Figure 3.2. We used the algorithm EG ($\eta = 0.05$) with the gradients of the ensemble CRPS. At the middle of the experiment, we swap the dispersion parameters of the members. Correct members become incorrect members and conversely. The members with incorrect dispersion parameter ($d_{ens} = 1.5$) see their weights decrease on average. After the swap, the weights of the newly incorrect members also decrease on average. The impact of the learning rate is shown in Figure 3.3, where a larger value $\eta = 0.2$ leads to a faster evolution of the weights. Note the difference of scales between Figures 3.2 and 3.3.

Now we show the average weight of the second subensemble parameterized by d_{ens}

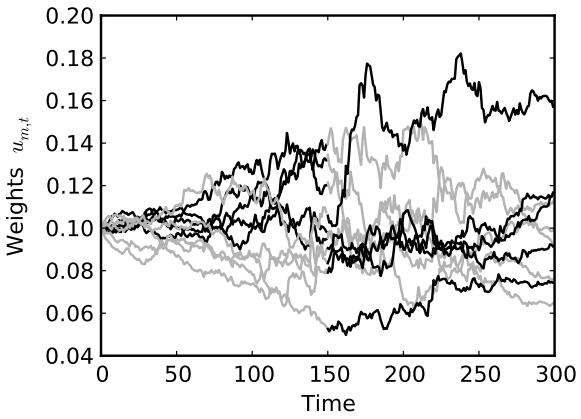


Figure 3.2 – Temporal evolution of the weights $u_{m,t}$, with learning rate $\eta = 0.05$. The weights of members with correct dispersion are in black, and the weights of members with the incorrect dispersion $d_{ens} = 1.5$ are in gray.

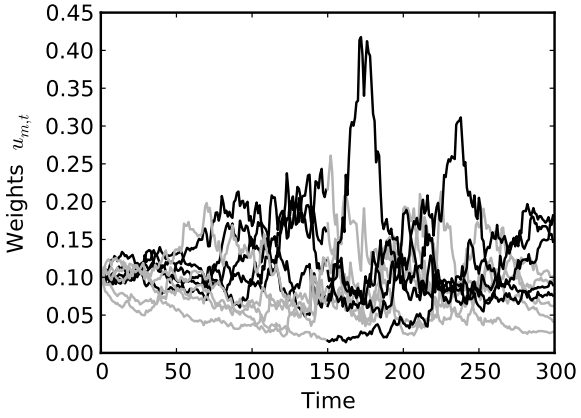


Figure 3.3 – Temporal evolution of the weights $u_{m,t}$, with learning rate $\eta = 0.2$. The weights of members with correct dispersion are in black, and the weights of members with the incorrect dispersion $d_{ens} = 1.5$ are in gray.

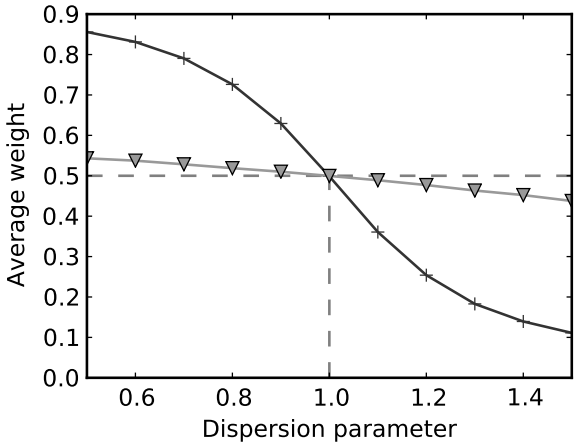


Figure 3.4 – Average cumulated weights of the members with (possibly) incorrect dispersion parameter d_{ens} (x-axis). The black crosses indicate that the CRPS of each member are used in EG (i). The light gray triangles indicate that the square loss gradients are used in EG (ii). This figure shows that not using the CRPS gradients favors the less dispersed members, even though they do not show the correct dispersion. The experiment of roughly ten years is repeated 200 times for each dispersion parameter. We used the learning parameters $\eta = 0.05$ and $\lambda = 0.5$.

for different learning algorithms. Here we did not include a change of the dispersion parameter at mid-experiment. The first subensemble therefore remains the correct one all the time. The discrimination procedure tests whether the algorithm makes a difference between the subensembles and whether the incorrect members are given lower weights than the correct members.

We show the importance of the CRPS gradients in EG for probabilistic forecasting. We show in Figure 3.4 the average weights of EG using: (i) $\ell_{m,t} = |x_{m,t} - y_t|$, using the CRPS without the gradients; or (ii) $\check{\ell}_{m,t} = 2(\boldsymbol{u}_t^\top \boldsymbol{x}_t - y_t)x_{m,t}$, using the gradients of the square loss $(\boldsymbol{u}_t^\top \boldsymbol{x}_t - y_t)^2$, instead of the CRPS gradients $\widetilde{\ell}_{m,t}$. We see that in either case, the members with the lowest dispersion parameter are the most weighted. The members with the correct distribution receive the highest weights when the incorrect members are over-dispersed ($d_{ens} > 1$). Formulation (i) and (ii) do not tend to forecast the distribution of the verifications, but only the mean or the median of the distribution of the verifications. These formulations are therefore not suited for probabilistic forecasting, as opposed to the CRPS gradients (see below). Note that we can rewrite $\check{\ell}_{m,t} = (x_{m,t} - y_t)^2 - (x_{m,t} - \boldsymbol{u}_t^\top \boldsymbol{x}_t)^2$ plus terms independent of m. Thus using the gradients (or equivalently trying to get the best combination) is a diversification strategy compared to simply using $(x_{m,t} - y_t)^2$.

Using the same representation, the algorithms EG and the ridge regression are tested with the ensemble CRPS and the class CRPS, see Figure 3.5. The algorithms based on the class CRPS show correct discrimination: whatever the dispersion parameter of the wrongly dispersed members, the class with incorrect dispersion shows smaller weights on average. The sum of the weights attributed to the incorrect members stays below 0.5 (equal weights between the two subensembles). On the contrary, the algorithms based

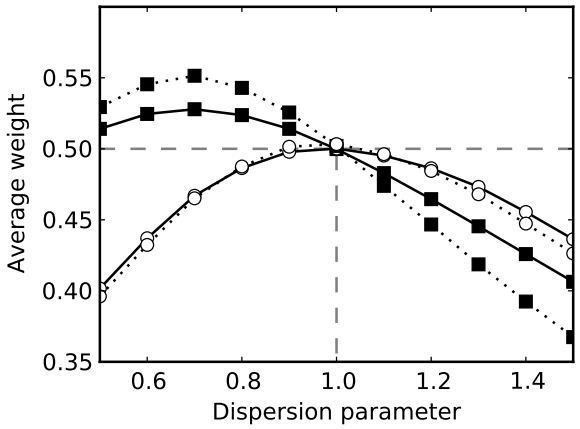


Figure 3.5 – Average cumulated weights of the members with (possibly) incorrect dispersion parameter. Both learning algorithms based on the CRPS are tested: EG (solid line) and ridge (dotted lined). The white circles indicate that the algorithm is run for class CRPS (equal weights within the class) and the black squares indicate that the weights are computed explicitly for each member.

on the ensemble CRPS does not give a correct discrimination. When the dispersion parameter d_{ens} is close to 0.70, the under-dispersed members receive larger weights than the correct members. We see that the minimization of the ensemble CRPS is misleading for an ensemble of small size. We interpret these results as direct consequences from the bias of the ensemble CRPS described in Section 3.1.4.

Conclusion

We introduced new tools for probabilistic forecasting using an ensemble of forecasts. Our algorithms use online learning techniques to produce forecast combinations that tend to minimize the CRPS. In the long run, they guarantee that the performance of the weighted ensemble is at least as good as the performance of the best weighted ensemble with weights constant in time. This theoretical guarantee holds with essentially no assumption on the distributions of the forecasts and verifications. In this sense, our method is a non-parametric post-processing method.

A new framework using classes of members is introduced in order to counteract the bias in the ensemble CRPS. With this framework and the proposed algorithms, numerical tests showed that our online learning techniques tend to give higher weights to the forecasts with the same distribution as the verifications.

The algorithms should now be tested against real data, in order to assess their potential in operational applications against Bayesian model averaging (BMA) or other post-processing techniques. The work of the forecaster is then to obtain numerous forecasts to combine. The methods do not require any assumptions on the forecasts to be applied (bias, spread, or any other stochastic or deterministic assumptions), except the loss boundedness. However, some good practices may be applied to improve the overall performance. For example, the forecasts can be altered before their inclusion

in the ensemble, or additional forecasts may be derived from the raw ensemble. Also, it is recommended to draw ensembles with enough spread, so that they encompass the verifications. We argue that for most applications, the use of a multimodel ensemble combined with several post-processing techniques is an efficient way to obtain an ensemble to be calibrated with our algorithms. From a meteorological point of view, new members can be added to the ensemble by using nearby grid-points or time-shifted forecasts. This approach may be particularly efficient to account for the ability of a forecasting system to predict an event, but at the wrong time or location.

On theoretical side, a next step could be the inclusion of the uncertainty in the verifications. Also, other non local strictly proper scoring rules could serve as loss function.

Appendix 3.A Identities implying CDFs

Let the random variable Z be described by the probability density function K' and the CDF K. We have for any real number x:

$$\mathrm{E}(\mathrm{H}(x-\mathrm{Z})) = \int \mathrm{K}'(\mathrm{Z})\mathrm{H}(x-\mathrm{Z})\mathrm{d}\mathrm{Z} = \mathrm{K}(x)\,, \tag{3.28}$$

or equivalently $\mathrm{E}(\mathrm{H}_{\mathrm{Z}}) = \mathrm{K}$. The demonstration of the strict propriety of the CRPS uses this property for the integration over the CDF of the verifications.

Let X and Z be two random variables described respectively by the CDFs G and K. We have:

$$\mathrm{E}(|\mathrm{X}-\mathrm{Z}|) = \int \mathrm{G}(1-\mathrm{K}) + \mathrm{K}(1-\mathrm{G})\,. \tag{3.29}$$

For $\mathrm{G} = \mathrm{K}$, the above quantity is the Gini mean difference, which is thoroughly introduced in the monograph of Yitzhaki and Schechtman [YS12].

The product GK of CDFs is itself the CDF of the random variable $\max(\mathrm{X}, \mathrm{Z})$. This can be used to explain simply Equation 3.29, using:

$$2\max(a,b) = |a-b| + a + b\,, \tag{3.30}$$

for any $(a,b) \in \mathbb{R}^2$, and

$$\mathrm{E}(\mathrm{Z}) = \int_{-\infty}^{+\infty} (\mathrm{H}(x) - \mathrm{K}(x))\mathrm{d}x\,. \tag{3.31}$$

Let $\mathrm{G} = \sum_{i=1}^{I} u_i \mathrm{G}_i$ and $\mathrm{K} = \sum_{j=1}^{J} w_j \mathrm{K}_j$ be CDFs of mixture models with respectively I and J components, i.e., the G_i and K_j are CDFs, and the weight vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ respectively belong to the simplexes $\mathcal{P}_I$ and $\mathcal{P}_J$. Let X, X_i, Z and Z_j be random variables respectively following G, G_i, K, and K_j. We have

$$\mathrm{E}(|\mathrm{X}-\mathrm{Z}|) = \sum_{i=1}^{I}\sum_{j=1}^{J} u_i w_j\, \mathrm{E}(|\mathrm{X}_i - \mathrm{Z}_j|)\,, \tag{3.32}$$

based on Equation 3.29. Indeed,

$$\sum_{i=1}^{I}\sum_{j=1}^{J} u_i w_j \int (\mathrm{G}_i(1-\mathrm{K}_j)+\mathrm{K}_j(1-\mathrm{G}_i)) = \int \sum_{i=1}^{I} u_i \mathrm{G}_i(1-\sum_{j=1}^{J} w_j \mathrm{K}_j) + \sum_{j=1}^{J} w_j \mathrm{K}_j(1-\sum_{i=1}^{I} u_i \mathrm{G}_i)$$
$$= \mathrm{E}(|\mathrm{X}-\mathrm{Z}|)\,,$$

because the weights w_i and u_j respectively sum to one, and using the linearity of integration.

It is straightforward to use Equation 3.32 to show to that

$$\mathrm{E}\left(|\mathrm{X}-y|\right) = \sum_{i=1}^{M} u_i \,\mathrm{E}\left(|\mathrm{X}_i - y|\right)\,, \tag{3.33}$$

and that

$$\mathrm{E}\left(|\mathrm{X}-\mathrm{X}'|\right) = \sum_{i=1}^{M} u_i \,\mathrm{E}\left(|\mathrm{X}_i-\mathrm{X}|\right) = \sum_{i,j=1}^{M} u_i u_j \,\mathrm{E}\left(|\mathrm{X}_i-\mathrm{X}'_j|\right)\,, \tag{3.34}$$

with X' and X'_j being random variables respectively described by G and G_j.

Appendix 3.B Computation of the ensemble CRPS

We have

$$\begin{aligned}
\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) = \int \Bigg(& \sum_{m,k=1}^{M} u_m u_k \mathrm{H}(x-x_m)\mathrm{H}(x-x_k) \\
& -2\sum_{m=1}^{M} u_m \mathrm{H}(x-x_m)\mathrm{H}(x-y) + \mathrm{H}(x-y)\Bigg)\,\mathrm{d}x \\
= & \sum_{m,k=1}^{M} u_m u_k(\Gamma - \max(x_m,x_k)) - 2\sum_{m=1}^{M} u_m(\Gamma - \max(x_m,y)) + \Gamma - y \\
= & -\sum_{m,k=1}^{M} u_m u_k \max(x_m,x_k) + 2\sum_{m=1}^{M} u_m \max(x_m,y) - y\,,
\end{aligned}$$

where Γ is the upper bound of the integral. Because the weights sum to one, we get the last simplification.

We rewrite the above expression using Equation 3.30:

$$\begin{aligned}
\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) = & -\frac{1}{2}\left(\sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + 2\sum_{m=1}^{M} u_m x_m\right) \\
& + \sum_{m=1}^{M} u_m |x_m - y| + \sum_{m=1}^{M} u_m(x_m+y) - y \\
= & \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2}\sum_{m,k=1}^{M} u_m u_k |x_m - x_k|\,,
\end{aligned} \tag{3.35}$$

because the weights u_m sum to one. We highlight the fact that the diagonal terms $u_m^2|x_m - x_m|$ are null, so that the double sum of Equation 3.35 is computed for $m \neq k$.

The calculus of this section can also be written with expectations and random variables using the content of Appendix 3.A.

Appendix 3.C Regret bound of the ridge regression with the CRPS

This section is written for general model mixtures $\mathrm{G}m,t$ and for general CDF F_t for the verifications. For simplicity, we assume that the integrals of the CRPS can be computed on an interval $[\gamma, \Gamma]$ of limited size. All the considered CDFs hit 0 at γ and 1 at Γ, which formalizes the assumption of bounded values for the members and the verifications. Thus the considered CDFs verify $\int G_{m,t} \leqslant \Gamma - \gamma$.

This appendix is structured as follows: (i) we exhibit an update rule between $\boldsymbol{u}_{t+1}$ and $\boldsymbol{u}_t$; (ii) we bound the regret against a constant vector $\sum_{t=1}^T \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{u})$ by the regret against the best a posteriori vector $\sum_{t=1}^T \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{u}_{t+1})$; (iii) we provide an interpretable regret bound by using the update rule and the convexity of ℓ_t.

The CRPS has a quadratic form

$$\ell_t(\boldsymbol{u}) = \boldsymbol{u}^\top \left(\int \mathbf{G}_t \mathbf{G}_t^\top \right) \boldsymbol{u} - 2\boldsymbol{u}^\top \int \mathrm{F}_t \mathbf{G}_t + \int \mathrm{F}_t^2 \,, \tag{3.36}$$

where $\mathrm{F}_t(x) = \mathrm{H}(x - y_t)$ and $\mathbf{G}_t(x)$ is the vector of the CDFs $\mathrm{G}_{m,t}(x)$.

The cost function $J_t(\boldsymbol{u}) = \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t'=1}^t \ell_{t'}(\boldsymbol{u})$ is written in a quadratic matricial form with:

$$J_t(\boldsymbol{u}) = \boldsymbol{u}^\top \mathbf{A}_{t+1} \boldsymbol{u} - 2\boldsymbol{u}^\top \boldsymbol{b}_{t+1} + \sum_{t'=1}^{t} \int \mathrm{F}_{t'}^2 \,, \tag{3.37}$$

where the vector $\boldsymbol{b}_t$ is defined by:

$$\boldsymbol{b}_t = \sum_{t'=1}^{t-1} \int \mathrm{F}_{t'} \mathbf{G}_{t'} \,, \tag{3.38}$$

and the matrix $\mathbf{A}_t$ of size $M \times M$ is symmetrical positive-definite:

$$\mathbf{A}_t = \lambda \mathbf{I}_M + \sum_{t'=1}^{t-1} \int \mathbf{G}_{t'} \mathbf{G}_{t'}^\top \,, \tag{3.39}$$

with $\mathbf{I}_M$ the identity matrix. The matrix $\mathbf{A}_t$ admits an inverse which is also symmetrical positive-definite. Note the trivial recurrence relation $J_{t+1} = \ell_{t+1} + J_t$.

The weight $\boldsymbol{u}_{t+1}$ is by definition the minimizer of J_t. Simple derivation gives the equality $\mathbf{A}_t \boldsymbol{u}_t = \boldsymbol{b}_t$. In practice, the weights are found via matrix inversion. Besides, a recurrence relation can be obtained. We successively deduce:

$$\begin{aligned}
\mathbf{A}_{t+1} \boldsymbol{u}_{t+1} = \boldsymbol{b}_{t+1} &= \boldsymbol{b}_t + \int \mathrm{F}_t \mathbf{G}_t \,, \\
&= \mathbf{A}_t \boldsymbol{u}_t + \int \mathrm{F}_t \mathbf{G}_t \,, \\
&= \left(\mathbf{A}_{t+1} - \int \mathbf{G}_t \mathbf{G}_t^\top \right) \boldsymbol{u}_t + \int \mathrm{F}_t \mathbf{G}_t \,.
\end{aligned}$$

The recurrence relation holds for any quadratic definition of the loss ℓ, and is expressed as:

$$\begin{aligned}\boldsymbol{u}_{t+1} - \boldsymbol{u}_t &= \mathbf{A}_{t+1}^{-1} \int \left(\mathrm{F}_t - \boldsymbol{u}_t^\top \mathbf{G}_t\right) \mathbf{G}_t \\ &= -\frac{1}{2}\mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t)\,. \end{aligned} \tag{3.40}$$

Demonstration of the regret bound

We iteratively use the fact that $\boldsymbol{u}_{t+1}$ is the minimizer of J_t to get

$$\begin{aligned} J_T(\boldsymbol{u}) &\geqslant J_T(\boldsymbol{u}_{T+1}) = \ell_T(\boldsymbol{u}_{T+1}) + J_{T-1}(\boldsymbol{u}_{T+1}) \\ &\geqslant \ell_T(\boldsymbol{u}_{T+1}) + J_{T-1}(\boldsymbol{u}_T) \\ &\geqslant \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_{t+1}) + \lambda \boldsymbol{u}_1^\top \boldsymbol{u}_1\,. \end{aligned} \tag{3.41}$$

The nonnegativity of $\lambda \boldsymbol{u}_1^\top \boldsymbol{u}_1$ gives:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \geqslant \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_{t+1}) - \lambda \boldsymbol{u}^\top \boldsymbol{u}\,. \tag{3.42}$$

Thus the regret can be bounded:

$$\begin{aligned} \mathcal{R}_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \\ &\leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}_{t+1}) \\ &\leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t=1}^{T} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top (\boldsymbol{u}_t - \boldsymbol{u}_{t+1}) \\ &= \lambda \boldsymbol{u}^\top \boldsymbol{u} + \frac{1}{2} \sum_{t=1}^{T} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top \mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t)\,, \end{aligned} \tag{3.43}$$

where we have used Equation 3.42, the convexity of the functions ℓ_t and Equation 3.40. At this point of the demonstration, one may have the feeling that a logarithm bound can be obtained, because the matrix $\mathbf{A}_t$ is a sum of t matrices, and because the logarithmic function is the primitive of the inverse function.

We define $\mathbf{Q}_t = \mathbf{A}_{t+1}^{-1/2} \mathbf{G}_t$ and $s_t = (\boldsymbol{u}_t^\top \mathbf{G}_t - \mathrm{F}_t)$, so that the symmetry of $\mathbf{A}_{t+1}^{-1/2}$ gives

$$\frac{1}{2} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top \mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t) = 2 \left(\int s_t \mathbf{Q}_t\right)^\top \left(\int s_t \mathbf{Q}_t\right)\,.$$

The inequality of Cauchy-Schwartz gives

$$\begin{aligned}\left(\int s_t \mathbf{Q}_t\right)^\top \left(\int s_t \mathbf{Q}_t\right) &= \sum_{m=1}^{M} \left[\left(\int s_t \mathbf{Q}_t\right)_m\right]^2 \\ &\leqslant \sum_{m=1}^{M} \int s_t^2 \int [(\mathbf{Q}_t)_m]^2 \\ &= \int s_t^2 \left(\int \mathbf{Q}_t^\top \mathbf{Q}_t\right) \\ &= \ell_t(\boldsymbol{u}_t) \left(\int \mathbf{G}_t^\top \mathbf{A}_{t+1}^{-1} \mathbf{G}_t\right) . \end{aligned} \tag{3.44}$$

We continue with

$$\begin{aligned}\int \mathbf{G}_t^\top \mathbf{A}_{t+1}^{-1} \mathbf{G}_t &= \operatorname{Tr}\left(\int \mathbf{A}_{t+1}^{-1} \mathbf{G}_t \mathbf{G}_t^\top\right) \\ &= \operatorname{Tr}\left(\mathbf{A}_{t+1}^{-1} \int \mathbf{G}_t \mathbf{G}_t^\top\right) \\ &= \operatorname{Tr}\left(\mathbf{I}_M - \mathbf{A}_{t+1}^{-1} \mathbf{A}_t\right) \\ &\leqslant \ln \frac{\det \mathbf{A}_{t+1}}{\det \mathbf{A}_t} . \end{aligned} \tag{3.45}$$

The first equality holds with the linearity of the integration and because $\boldsymbol{z}_1^\top \mathbf{A} \boldsymbol{z}_2 = \operatorname{Tr}(\mathbf{A} \boldsymbol{z}_2 \boldsymbol{z}_1^\top)$ for any vectors $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and matrix $\mathbf{A}$. The inequality holds because $\mathbf{A}_{t+1}^{-1} \mathbf{A}_t$ is positive definite and $1 - 1/x \leqslant \ln x$ for any $x > 0$.

At this step of the proof, we have shown that:

$$\mathcal{R}_T(\boldsymbol{u}) \leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2 \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \ln \frac{\det \mathbf{A}_{t+1}}{\det \mathbf{A}_t} . \tag{3.46}$$

We assume that the losses $\ell_t(\boldsymbol{u}_t)$ are bounded by $a > 0$. Then we easily reach:

$$\mathcal{R}_T(\boldsymbol{u}) \leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2a \ln \frac{\det \mathbf{A}_{T+1}}{\lambda^M} . \tag{3.47}$$

The inequality of arithmetic and geometric means applied to the eigenvalues of $\mathbf{A}_{T+1}$ leads to the conclusion

$$\begin{aligned}\det(\mathbf{A}_{T+1}) &\leqslant \left(\frac{\operatorname{Tr} \mathbf{A}_{T+1}}{M}\right)^M = \left(\frac{M\lambda + \sum_{t=1}^{T} \operatorname{Tr} \int \mathbf{G}_t \mathbf{G}_t^\top}{M}\right)^M \\ &\leqslant \left(\frac{M\lambda + MT(\Gamma - \gamma)}{M}\right)^M , \end{aligned} \tag{3.48}$$

from which we conclude that

$$\begin{aligned}\mathcal{R}_T(\boldsymbol{u}) &\leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2aM \ln \left(1 + \frac{T(\Gamma - \gamma)}{\lambda}\right) \\ &\leqslant \lambda \boldsymbol{u}^\top \boldsymbol{u} + \mathcal{O}(\ln T) . \end{aligned} \tag{3.49}$$

We logically compete against any constant vector $\boldsymbol{u}$ on the simplex so that

$$\sup_{\boldsymbol{u}\in\mathcal{P}_M} \mathcal{R}_T(\boldsymbol{u}) \leqslant \mathcal{O}(\ln T)\,. \tag{3.50}$$

□

4 Scoring and learning forecasts densities

The purpose of this chapter is the generalization of the results of Chapter 3. In a first part, we generalize our results to non-local strictly proper scoring rules, other than the CRPS. We focus our attention on scoring rules admitting a threshold or a quantile decomposition. Relationships between this decomposition, score biases and model mixtures are investigated. The question of noisy observations is addressed in the second part of this chapter. We include uncertainty information in the CRPS and apply a generalized least-square procedure. Interestingly, the expectation of the derived loss is related to Pearson's χ^2 statistic.

Contents

4.1 Extension to threshold-weighted and quantile-weighted scoring rules

As introduced in Section 1.3, various ways exist to evaluate forecasts of binary events. A infinite amount of strictly proper scoring rules can be written as a sum of elementary quantile losses. Besides, a second summation on event thresholds allows to evaluate probabilistic forecasts of a scalar variable. In this section, the notation α refers to levels of quantile and the notation θ refers to thresholds. The interested reader is referred to the work of Buja et al. [BSS05], Gneiting and Ranjan [GR11], and Ehm et al. [Ehm+16] for state-of-the-art articles on this subject, and Dawid [Daw08] for a detailed bibliographical analysis of assessments of probabilistic forecasts. General conditions ensuring strict propriety are not new since they were addressed by Shuford et al. [SAE66], Savage [Sav71], and Schervish [Sch89], among others. The choice between these scoring rules was also recently addressed by Merkle and Steyvers [MS13] and Lerch et al. [Ler+15].

In the remaining of this short introduction, we briefly review the score decomposition of Ehm et al. [Ehm+16] for the celebrated CRPS. In Section 4.1.1, we generalize the results of Thorey et al. [TMB16] to threshold-weighted scoring rules. Quantile-weighted scoring rules are analyzed in the context of mixture models in Section 4.1.2.

Let G be a CDF delivered to forecast the observation y. The CRPS can be written

$$\mathrm{CRPS}(\mathrm{G}, y) = 2 \iint S_{\alpha,\theta}(\mathrm{G}^{-1}(\alpha), y) \mathrm{d}\alpha \mathrm{d}\theta \,, \tag{4.1}$$

if G is an invertible function. The integration holds on the levels $\alpha \in [0, 1]$ of quantile and on the thresholds $\theta \in \mathbb{R}$. The function $S_{\alpha,\theta}(x, y)$ is defined by

$$\begin{aligned} S_{\alpha,\theta}(x, y) &= (\mathrm{H}(x - y) - \alpha)(\mathrm{H}(x - \theta) - \mathrm{H}(y - \theta)) \\ &= \begin{cases} 1 - \alpha & \text{if } y \leqslant \theta < x \,, \\ \alpha & \text{if } x \leqslant \theta < y \,, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.2}$$

The forecast x of the observation y is evaluated with the score $S_{\alpha,\theta}(x, y)$ for the event $y < \theta$. The quantity α is the cost of a false positive, and $1 - \alpha$ is the cost a false negative.

Integrating over the levels α gives the standard formulation of the CRPS, indeed

$$\begin{aligned} 2 \int S_{\alpha,\theta}(\mathrm{G}^{-1}(\alpha), y) \mathrm{d}\alpha &= \begin{cases} 2 \int_\theta (1 - \mathrm{G}(x)) g(x) \mathrm{d}x & \text{if } y \leqslant \theta \,, \\ 2 \int^\theta \mathrm{G}(x) g(x) \mathrm{d}x & \text{if } \theta < y \,, \end{cases} \\ &= (\mathrm{H}(\theta - y) - \mathrm{G}(\theta))^2 \,, \end{aligned}$$

with the change of variable $\mathrm{G}(x) = \alpha$.

Integrating over the thresholds relates the CRPS to the quantile score $\mathrm{QS}_\alpha(x, y)$ (or pinball loss) of level α [KB78]. The quantile decomposition of the CRPS was found by Laio and Tamea [LT07], and developed by Bröcker [Brö12] for ensemble forecasts. Using $\int (\mathrm{H}(x - \theta) - \mathrm{H}(y - \theta)) \mathrm{d}\theta = x - y$, we have

$$\int S_{\alpha,\theta}(x, y) \mathrm{d}\theta = (\mathrm{H}(x - y) - \alpha)(x - y) = \mathrm{QS}_\alpha(x, y) \,,$$

thus

$$\mathrm{CRPS}(\mathrm{G}, y) = 2 \int \mathrm{QS}_\alpha(\mathrm{G}^{-1}(\alpha), y) \mathrm{d}\alpha \,.$$

In the following, we elaborate on the impact of adding weighting functions $\phi(\theta)$ or $\omega(\alpha)$ in Equation 4.1 instead of uniform weighting schemes.

4.1.1 Effect of threshold-weighting

Here, we show how our analysis on the bias of the ensemble CRPS and on the class CRPS is easily extended to the threshold-weighted CRPS. The results of Section 3.1 hold up to a change of variables, suggested by Ehm et al. [Ehm+16].

The CRPS, seen as a sum of Brier scores can be extended to the threshold-weighted CRPS [MW76], where different weights are used for each threshold. Using the strictly positive function ϕ, the threshold-weighted CRPS is defined by:

$$\mathrm{wCRPS}(\mathrm{G}, y) = \int (\mathrm{G} - \mathrm{H}_y)^2 \, \phi \,. \tag{4.3}$$

Let Φ be the antiderivative function of ϕ, the function Φ is strictly increasing and invertible because the function ϕ is strictly positive.

We now get into further details by generalizing Equation 3.29 and Equation 3.31 of Appendix 3.A. Let X and Z be independent random variables respectively described by the CDFs G and K, and let H be the unit step function centered on 0. We have

$$\int (\mathrm{H} - \mathrm{K}) \phi = \mathrm{E}(\Phi(\mathrm{Z})) - \Phi(0) \,, \tag{4.4}$$

using integration by parts, and, as proved below,

$$\mathrm{E}(|\Phi(\mathrm{X}) - \Phi(\mathrm{Z})|) = \int \phi(\mathrm{G} + \mathrm{K} - 2\mathrm{GK}) \,. \tag{4.5}$$

We see that weighting thresholds with ϕ is equivalent to a data transformation with Φ. We now demonstrate Equation 4.5:

$$\begin{aligned}
\int \phi(\mathrm{G} + \mathrm{K} - 2\mathrm{GK}) &= \int \phi(\mathrm{G} + \mathrm{K} - 2\mathrm{H} + 2\mathrm{H} - 2\mathrm{GK}) \\
&= 2\,\mathrm{E}(\Phi(\max(\mathrm{X}, \mathrm{Z}))) - \mathrm{E}(\Phi(\mathrm{X})) - \mathrm{E}(\Phi(\mathrm{Z})) \\
&= 2\,\mathrm{E}(\max(\Phi(\mathrm{X}), \Phi(\mathrm{Z}))) - \mathrm{E}(\Phi(\mathrm{X})) - \mathrm{E}(\Phi(\mathrm{Z})) \\
&= \mathrm{E}(|\Phi(\mathrm{X}) - \Phi(\mathrm{Z})|) \,.
\end{aligned}$$

We have successively used the fact that GK is the CDF of $\max(\mathrm{X}, \mathrm{Z})$, Equation 4.4, the fact that for any $x, z \in \mathbb{R}$, $\Phi(\max(x, z)) = \max(\Phi(x), \Phi(z))$ because Φ is an increasing function, and $2\max(x, z) = |x - z| + x + z$.

With the same notation as in Section 3.1.4, we show that the bias of the ensemble threshold-weighted CRPS with underlying mixture model is therefore similar to the unweighted case of Equation 3.7 up to the data transformation with Φ. Let $\mathrm{G}^{\mathcal{E}}$ be the weighted CDF described by the members x_m and the weights u_m. The members are assumed to be samples from the independent random variables X_m. The CDF

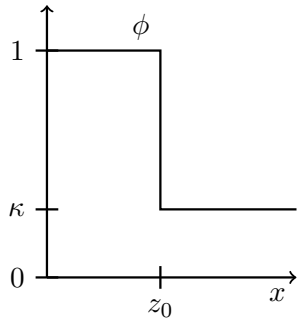


Figure 4.1 – Illustration of the step function ϕ.

$\mathrm{G} = \mathrm{E}(\mathrm{G}^{\mathcal{E}})$, where the expectation is taken over the members, describes the random variable X. With Equation 4.5, we swap the members x_m for $\Phi(x_m)$ and the verification y for $\Phi(y)$ in the expression of the weighted CRPS:

$$\mathrm{wCRPS}(\mathrm{G}^{\mathcal{E}}, y) = \sum_{m=1}^{M} u_m |\Phi(x_m) - \Phi(y)| - \frac{1}{2} \sum_{m,k=1}^{M} u_m u_k |\Phi(x_m) - \Phi(x_k)| \,. \tag{4.6}$$

Therefore, the bias of the threshold-weighted CRPS is explicitly found in

$$\begin{aligned} \mathrm{E}(\mathrm{wCRPS}(\mathrm{G}^{\mathcal{E}}, y)) &= \mathrm{E}(|\Phi(\mathrm{X}) - \Phi(y)|) - \frac{1}{2}\,\mathrm{E}(|\Phi(\mathrm{X}) - \Phi(\mathrm{X}')|) \\ &\quad + \frac{1}{2} \sum_{m=1}^{M} u_m^2\,\mathrm{E}(|\Phi(\mathrm{X}_m) - \Phi(\mathrm{X}'_m)|) \\ &= \mathrm{wCRPS}(\mathrm{G}, y) + \frac{1}{2} \sum_{m=1}^{M} u_m^2\,\mathrm{E}(|\Phi(\mathrm{X}_m) - \Phi(\mathrm{X}'_m)|) \,. \end{aligned}$$

Also for the class CRPS, our results are extended to the class threshold-weighted CRPS by introducing the quantities of interest $\sum_{c=1}^{M_C} |\Phi(x_c^C) - \Phi(y)|/M_C$, $\sum_{c=1}^{M_C} \sum_{d=1}^{M_D} |\Phi(x_c^C) - \Phi(x_d^D)|/(M_C M_D)$, and $\sum_{c,c'=1}^{M_C} |\Phi(x_c^C) - \Phi(x_{c'}^C)|/(M_C(M_C - 1))$.

In order to give a better understanding on the effect of the Φ-transformation, we give the following example. Consider the case of a step function ϕ defined by:

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leqslant x < z_0 \\ \kappa & \text{if } z_0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases},$$

with $\kappa < 1$, $z_0 \in [0, 1]$ (see Figure 4.1). We assume that all observations and members are in $[0, 1]$. Roughly speaking, in our example, errors in $[0, z_0]$ cost more than errors in $[z_0, 1]$. In this setting, we have

$$\Phi(x) = \begin{cases} x & \text{if } 0 \leqslant x < z_0 \\ z_0 + \kappa(x - z_0) & \text{if } z_0 < x \leqslant 1 \,, \end{cases}$$

and

$$|\Phi(x_m) - \Phi(x_k)| = \begin{cases} |x_m - x_k| & \text{if } 0 \leqslant x_k, x_m \leqslant z_0 \\ \kappa|x_m - x_k| & \text{if } z_0 \leqslant x_k, x_m \leqslant 1 \\ |z_0 - x_m| + \kappa|x_k - z_0| & \text{if } 0 \leqslant x_m \leqslant z_0 \leqslant x_k \leqslant 1\,. \end{cases}$$

In this simple example, it is easy to see that the distance $|\Phi(x_m) - \Phi(x_k)|$ is high when x_m and x_k are in the domain of high value of ϕ. And conversely, the distance $|\Phi(x_m) - \Phi(x_k)|$ is low when x_m and x_k are in the domain of low value of ϕ. This is consistent with the intuition that the function ϕ defines domains with high or low importance.

4.1.2 Effect of quantile-weighting

We prove in this section that quantile weighting can also be understood as a data transformation, at least for the score gradients. We show that the data transformation due to the quantile weighting is related to both the weighting function ω and the forecaster's CDF G. We begin this section by rewriting quantile-weighted scores in a convenient way and studying the quantile-weighted score of an ensemble of forecasts.

Rewriting quantile-weighted scoring rules

Let the score $S(\mathrm{G}, y)$ be a quantile-weighted score defined by

$$\begin{aligned} S(\mathrm{G}, y) &= \int \mathrm{QS}_\alpha(\mathrm{G}^{-1}(\alpha), y)\omega(\alpha)\mathrm{d}\alpha \\ &= \int \mathrm{QS}_{\mathrm{G}(\mathrm{X})}(\mathrm{X}, y)\omega(\mathrm{G}(\mathrm{X}))\mathrm{dG}(\mathrm{X}) \\ &= \mathrm{E}_\mathrm{X}\left[\mathrm{QS}_{\mathrm{G}(\mathrm{X})}(\mathrm{X}, y)\omega(\mathrm{G}(\mathrm{X}))\right], \end{aligned}$$

where ω is strictly positive on open intervals in $[0, 1]$. Propriety is retained since the minimizer of $\mathrm{E}_\mathrm{Y}[S(\mathrm{G}, \mathrm{Y})]$ verifies $\mathrm{G}(\alpha) = \mathrm{F}(\alpha)$ for each $0 < \alpha < 1$, where Y is a random variable described by the CDF F.

From the definition of $S_{\alpha,\theta}$ in Equation 4.2, we have

$$S(\mathrm{G}, y) = \int \left(\mathrm{H}(\theta - y)\left[\int_{\mathrm{G}(\theta)}^{1}(1-\alpha)\omega(\alpha)\mathrm{d}\alpha\right] + \mathrm{H}(y-\theta)\left[\int_0^{\mathrm{G}(\theta)}\alpha\omega(\alpha)\mathrm{d}\alpha\right]\right)\mathrm{d}\theta\,. \tag{4.7}$$

We used the conditions $y \leqslant \theta < x$ or equivalently $\mathrm{G}(y) \leqslant \mathrm{G}(\theta) < \alpha$ for the left term, and the conditions $x \leqslant \theta < y$ or equivalently $\alpha \leqslant \mathrm{G}(\theta) < \mathrm{G}(y)$ for the right term.

The score $S(\mathrm{G}, y)$ can therefore be expressed as

$$S(\mathrm{G}, y) = \int \mathrm{H}_y\beta_1(\mathrm{G}) + (1 - \mathrm{H}_y)\beta_0(1 - \mathrm{G})\,, \tag{4.8}$$

by identifying

$$\beta_1(\mathrm{G}) = \int_{\mathrm{G}}^{1}(1-\alpha)\omega(\alpha)\mathrm{d}\alpha \quad \text{and} \quad \beta_0(1-\mathrm{G}) = \int_0^{\mathrm{G}}\alpha\omega(\alpha)\mathrm{d}\alpha\,.$$

For example with the CRPS, we have $\beta_1(\mathrm{G}) = (1-\mathrm{G})^2$ and $\beta_0(1-\mathrm{G}) = \mathrm{G}^2$. If ω is symmetrical with respect to 1/2, then $\beta_1 = \beta_0$. Equation 4.7 proposes a formulation for quantile-weighted scoring rules emphasizing the interplay between ω and G. This formulation is particularly useful when analytical expressions of β_1 and β_0 are available.

The linearity in H_y is convenient to derive the uncertainty term of F and the divergence term between F and G in:

$$\begin{aligned}
\mathrm{E}_{\mathrm{Y}}[S(\mathrm{G},\mathrm{Y})] &= \int \mathrm{F}\beta_1(\mathrm{G}) + (1-\mathrm{F})\beta_0(1-\mathrm{G}) \\
&= \underbrace{\int \mathrm{F}(\beta_1(\mathrm{G}) - \beta_1(\mathrm{F})) + (1-\mathrm{F})(\beta_0(1-\mathrm{G}) - \beta_0(1-\mathrm{F}))}_{\text{divergence} = d(\mathrm{F},\mathrm{G})} \\
&\quad + \underbrace{\int \mathrm{F}\beta_1(\mathrm{F}) + (1-\mathrm{F})\beta_0(1-\mathrm{F})}_{\text{uncertainty} = e(\mathrm{F})} .
\end{aligned}$$

We highlight the fact that $\beta_1(\mathrm{G}) - \beta_1(\mathrm{F})$ may be rewritten under the form $\int_{\mathrm{G}}^{\mathrm{F}} (1-\alpha)\omega(\alpha)\mathrm{d}\alpha$ and similarly for $\beta_0(1-\mathrm{G}) - \beta_1(1-\mathrm{F})$. This gives an explicit formulation of the divergence term. Besides, the uncertainty term is equal to $\mathrm{E}_{\mathrm{Y}}[S(\mathrm{F},\mathrm{Y})]$.

By considering a distribution on the forecasts G, the above decomposition may be led further, as for any strictly proper scoring rule [Brö09]. To do so, we note $\overline{\mathrm{F}} = \mathrm{E}_{\mathrm{G}}[\mathrm{F}]$ the climatological distribution, where the expectation is taken with respect to the frequency at which G is predicted (F is implicitly conditioned on G). The reliability $\mathrm{E}_{\mathrm{G}}[d(\mathrm{F},\mathrm{G})]$ measures the match between the forecasted probabilities $\mathrm{G}(\theta)$ and the conditional frequencies $\mathrm{F}(\theta)$. The resolution $\mathrm{E}_{\mathrm{G}}[d(\mathrm{F},\overline{\mathrm{F}})]$ is the distance between the climatology $\overline{\mathrm{F}}(\theta)$ and the conditional frequencies $\mathrm{F}(\theta)$. Such separation dates back to Murphy [Mur73] for the Brier score, Hersbach [Her00] for the CRPS, and Bentzien and Friederichs [BF14] for quantile scores. The decomposition writes

$$\mathrm{E}_{\mathrm{G},\mathrm{Y}}[S(\mathrm{G},\mathrm{Y})] = \underbrace{e(\overline{\mathrm{F}})}_{\text{uncertainty}} + \underbrace{\mathrm{E}_{\mathrm{G}}[d(\mathrm{F},\mathrm{G})]}_{\text{reliability}} - \underbrace{\mathrm{E}_{\mathrm{G}}[d(\mathrm{F},\overline{\mathrm{F}})]}_{\text{resolution}} .$$

For any quantile-weighted scoring rule, the explicit formulations of the uncertainty e and the divergence d give the score decomposition into uncertainty, reliability and resolution.

We finish this section with examples of quantile-weighted scoring rules. Several weighting functions ω of the beta family including the CRPS are defined in Table 4.1, and represented in Figure 4.2. The consequence of the choice of ω is shown in Figure 4.3, which illustrates the functions $\mathrm{H}_y\beta_1(\mathrm{G}) + (1-\mathrm{H}_y)\beta_0(1-\mathrm{G})$ in two situations: $\mathrm{G}^{(1)}$ is the CDF of a Gaussian distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$, and $\mathrm{G}^{(2)}$ is the CDF of a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = (2.577)^2$. In both cases, a null observation $y = 0$ is considered. Higher importance is given to the distribution tails when ω assigns higher weights to the extreme quantile levels near 0 and 1. Conversely, higher importance is given to the distribution center when ω assigns higher weights to the quantile levels close to 0.5.

$\omega(\alpha)$	$\beta_1(\alpha)$	
$(\alpha(1-\alpha))^{-1}$	$-\ln(\alpha)$	(CRIGN)
$(\alpha(1-\alpha))^{-1/2}$	$\arcsin(\sqrt{1-\alpha}) - \sqrt{\alpha(1-\alpha)}$	
2	$(1-\alpha)^2$	(CRPS)
$\alpha(1-\alpha)$	$\frac{(1-\alpha)^3}{3} - \frac{(1-\alpha)^4}{4}$	

Table 4.1 – Examples of scoring rules in the symmetrical beta family $\omega(\alpha) = \alpha^{a-1}(1-\alpha)^{a-1}$.

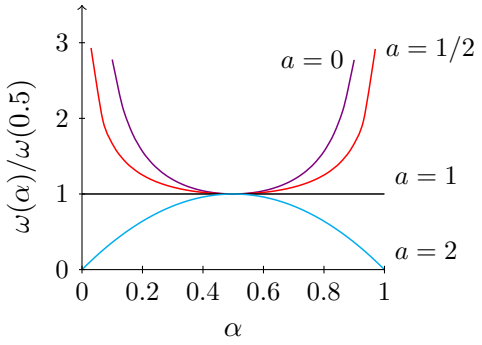


Figure 4.2 – Scaled weighting function $\omega/\omega(0.5)$ in the symmetrical beta family $\omega(\alpha) = \alpha^{a-1}(1-\alpha)^{a-1}$.

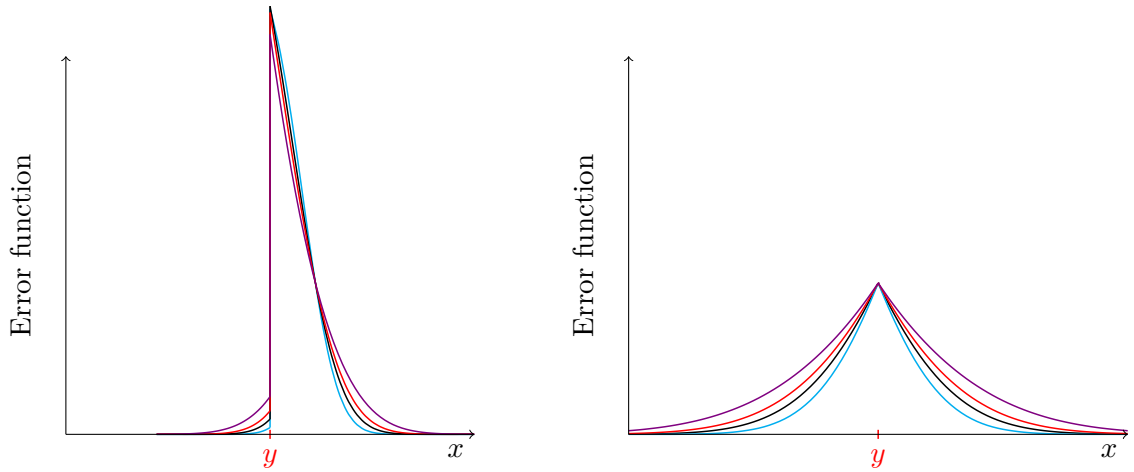


Figure 4.3 – Illustration of the error function $\mathrm{H}_y\beta_1(\mathrm{G}) + (1-\mathrm{H}_y)\beta_0(1-\mathrm{G})$ for $y = 0$ and G the CDF of a Gaussian distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$ (left), and $y = 0$ and G the CDF of a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = (2.577)^2$ (right). The weighting functions ω belong to the symmetrical beta family, $\omega(\alpha) = \alpha^{a-1}(1-\alpha)^{a-1}$: a = 0 (violet), a = 1/2 (red), a = 1 (black), a = 2 (cyan). The functions are scaled to reach the same value at $\mathrm{G}^{-1}(0.5)$.

Optimal quantiles locations for an ensemble of forecasts

In this section, we derive the optimal quantile positions of the M members x_m with fixed weights u_m when the combination $\mathrm{G}^{\mathcal{E}}$ of unit step functions is evaluated by a quantile-weighted scoring rule. Without lack of generality, we assume that the members are labeled according to their rank, i.e $x_{m-1} < x_m$ for any integer $m \in [2, M]$. We note $U_m = \sum_{k \leqslant m} u_k$ with the convention $U_0 = 0$. The optimal location of the members is well-known for the CRPS [Brö12]. The CRPS expectation over the observations is minimized by members verifying $\mathrm{F}(x_m) = U_m - \frac{u_m}{2}$. We show that the optimal locations for the CRPS are close to the optimal locations found for other quantile-weighted scoring rules, except for the outer members x_1 and x_M. For simplicity, the distribution of the observations is assumed to be bounded.

We recall that

$$\mathrm{E}_{\mathrm{Y}}[S(\mathrm{G}^{\mathcal{E}}, \mathrm{Y})] = \int \mathrm{F}\beta_1(\mathrm{G}^{\mathcal{E}}) + (1 - \mathrm{F})\beta_0(1 - \mathrm{G}^{\mathcal{E}})\,. \tag{4.9}$$

Since $\mathrm{G}^{\mathcal{E}}$ is piecewise constant, the term depending on x_m in $\mathrm{E}_{\mathrm{Y}}[S(\mathrm{G}^{\mathcal{E}}, \mathrm{Y})]$ is

$$\beta_1(U_{m-1}) \int_{x_{m-1}}^{x_m} \mathrm{F} + \beta_1(U_m) \int_{x_m}^{x_{m+1}} \mathrm{F} + \beta_0(1-U_{m-1}) \int_{x_{m-1}}^{x_m} (1-\mathrm{F}) + \beta_0(1-U_m) \int_{x_m}^{x_{m+1}} (1-\mathrm{F})\,, \tag{4.10}$$

for any integer $m \in [2, M-1]$. This expression is also valid for x_1 and x_M if β_1 and β_0 are bounded because $\beta_1(0) = \beta_0(0) = 0$. In this case, we use the convention $x_0 = \sup_{\mathrm{F}(x)=0} x$ and $x_{M+1} = \inf_{\mathrm{F}(x)=1} x$. If β_1 and β_0 are not bounded, the optimal location of the extreme members lies at the bounds of the observational distribution, i.e. $x_1 = \sup_{\mathrm{F}(x)=0} x$ and $x_M = \inf_{\mathrm{F}(x)=1} x$.

After differentiating Expression 4.10 with respect to x_m, and setting the derivative to 0, the optimal location of the members are found with

$$\mathrm{F}(x_m) = \frac{\beta_0(1 - U_m) - \beta_0(1 - U_{m-1})}{\beta_0(1 - U_m) - \beta_0(1 - U_{m-1}) + \beta_1(U_{m-1}) - \beta_1(U_m)}\,. \tag{4.11}$$

A safety check for the CRPS indicates that Equation 4.11 is consistent with the results of Bröcker [Brö12]:

$$\begin{aligned}
\mathrm{F}(x_m) &= \frac{U_m^2 - (U_m - u_m)^2}{U_m^2 - (U_m - u_m)^2 + (1 - U_m + u_m)^2 - (1 - U_m)^2} \\
&= \frac{u_m}{u_m} \times \frac{2U_m - u_m}{2U_m - u_m + 2 - 2U_m + u_m} \\
&= U_m - \frac{u_m}{2}\,,
\end{aligned}$$

with $\beta_1(\alpha) = (1 - \alpha)^2$ and $\beta_0(1 - \alpha) = \alpha^2$.

The optimal quantiles of several scoring rules are shown in Figure 4.4. The location of the optimal members do not vary much depending on the scoring rule, except for the position of x_1 and x_M. Larger variations are observed for a small amount of members M. The location of the members are barely distinguishable for $M = 20$.

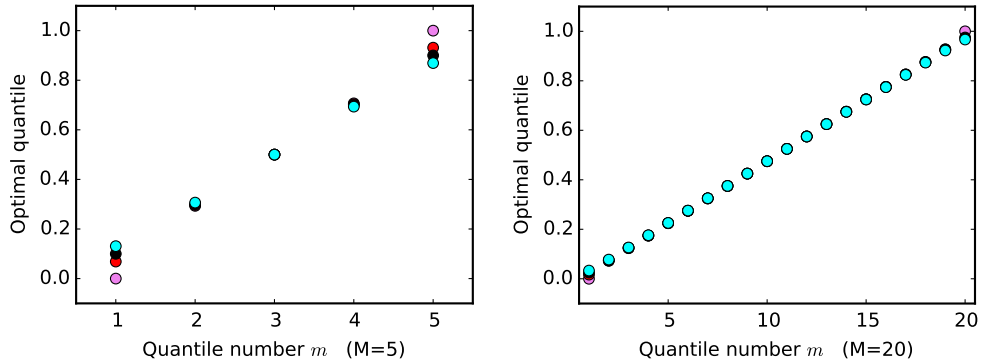


Figure 4.4 – Optimal quantile levels $\mathrm{F}(x_m)$ of the members for $M = 5$ (left) and $M = 20$ (right) with weighting functions ω belonging to the symmetrical beta family, $\omega(\alpha) = \alpha^{a-1}(1-\alpha)^{a-1}$: a = 0 (violet), a = 1/2 (red), a = 1 (black), a = 2 (cyan).

Model mixture and quantile-weighted scoring rules

Now we investigate the relationships between quantile-weighted scores and their gradients for model mixtures. Let $\mathrm{G} = \sum_{m \leqslant M} u_m \mathrm{G}_m$ be a model mixture. We compute the gradients of $S(\mathrm{G}, y)$ with Equation 4.7:

$$\begin{aligned}\frac{\partial S(\mathrm{G}, y)}{\partial u_m} &= \int \mathrm{H}_y[-\mathrm{G}_m(1-\mathrm{G})\omega(\mathrm{G})] + (1-\mathrm{H}_y)\mathrm{G}_m\mathrm{G}\omega(\mathrm{G}) \\ &= \int \mathrm{G}_m(\mathrm{G}-\mathrm{H}_y)\omega(\mathrm{G}) \\ &= 1/2 \int [(\mathrm{H}_y + \mathrm{G}_m - 2\mathrm{H}_y\mathrm{G}_m) - (\mathrm{G} + \mathrm{G}_m - 2\mathrm{G}\mathrm{G}_m) + \mathrm{G} - \mathrm{H}_y]\omega(\mathrm{G})\,.\end{aligned} \tag{4.12}$$

We check that we find the CRPS for $\omega = 2$. In the Expression 4.12 of $\frac{\partial S}{\partial u_m}$, we find a balance between how close G_m is to H_y against how close G_m is to G. The distance is measured by terms of the form $\int(\mathrm{K}_1 + \mathrm{K}_2 - 2\mathrm{K}_1\mathrm{K}_2)\omega(\mathrm{G})$, which can be seen as a generalization of the Gini mean difference with the additional $\omega(\mathrm{G})$ term (see Equation 4.5). Similarly to the CRPS, the distance between G_m and G is a weighted distance between the mth members and the other members. Indeed we have

$$\int(\mathrm{G} + \mathrm{G}_m - 2\mathrm{G}\mathrm{G}_m)\omega(\mathrm{G}) = \sum_{k=1}^{M} u_k \int(\mathrm{G}_k + \mathrm{G}_m - 2\mathrm{G}_k\mathrm{G}_m)\omega(\mathrm{G})\,. \tag{4.13}$$

The quantile weighting can therefore be understood as a data transformation for the score gradients, recalling the results obtained for threshold-weighted scoring rules in Section 4.1.1. For quantile-weighted scoring rules, the weighting $\phi = \omega(\mathrm{G})$ is applied in Equation 4.12 to the scores gradients. The data transformation related to the weighting $\omega(\mathrm{G})$ is its antiderivative $\int^x \omega(\mathrm{G})$. Note that this data transformation depends on G.

To conclude this section, we study the effect of the quantile-weighting for ensemble forecasting, where $\mathrm{G}^{\mathcal{E}}$ and $\omega(\mathrm{G}^{\mathcal{E}})$ are piecewise constant functions. This case is close to

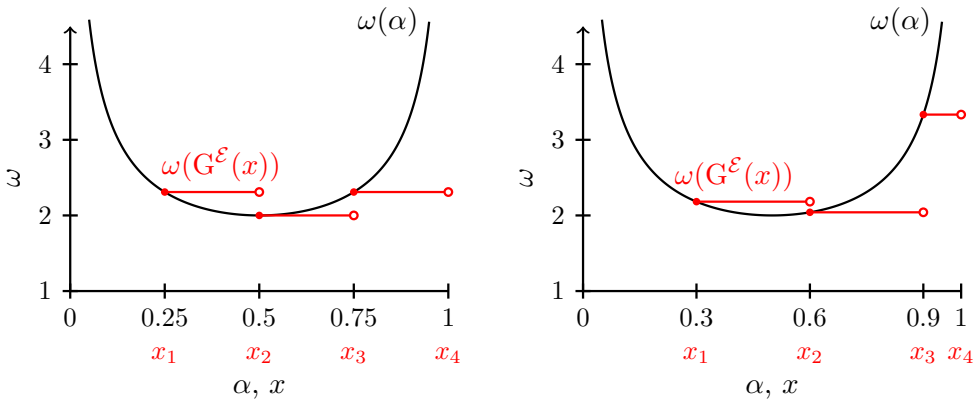


Figure 4.5 – Illustration of the quantile-weighting function ω (black), for a convex function $\omega = \frac{1}{\sqrt{\alpha(1-\alpha)}}$. The members x_m and the weighting function $\omega(\mathrm{G}^{\mathcal{E}})$ (red) are shown to illustrate the importance of the intervals $[x_m, x_{m+1}]$ in two cases: uniform weights $u_1 = u_2 = u_3 = u_4 = 0.25$ (left) and non uniform weights $u_1 = 0.3$, $u_2 = 0.3$, $u_3 = 0.3$, $u_4 = 0.1$ (right).

the example provided for threshold-weighted score in Section 4.1.1, with the piecewise constant weighting function ϕ.

Let $(z, z') \in \mathbb{R}^2$ play the role of (x_m, y) or (x_m, x_k). We have

$$\int (\mathrm{H}(x-z) + \mathrm{H}(x-z') - 2\mathrm{H}(x-z)\mathrm{H}(x-z'))\omega(\mathrm{G}^{\mathcal{E}}(x))\mathrm{d}x = |\int_z^{z'} \omega(\mathrm{G}^{\mathcal{E}})|\,, \quad (4.14)$$

because the quantity $(\mathrm{H}(x-z) + \mathrm{H}(x-z') - 2\mathrm{H}(x-z)\mathrm{H}(x-z')) = 1$ if x is between z and z', and zero otherwise. Recalling that $\omega(\mathrm{G}^{\mathcal{E}})$ is piecewise constant, we see that the gradient $\frac{\partial S(\mathrm{G}^{\mathcal{E}}, y)}{\partial u_m}$ is controlled by terms $|\int_z^{z'} \omega(\mathrm{G}^{\mathcal{E}})|$ which are piecewise affine in z and z'. The slope $\omega(\mathrm{G}^{\mathcal{E}}(x))$, with $x \in [x_m, x_{m+1}]$, can be seen as the importance given to the interval $[x_m, x_{m+1}]$.

In practice, the gradients may be computed as a sum of terms $(x_{m+1} - x_m)\omega(U_m)$, accompanied by $(x_{m^*+1} - y)\omega(U_{m^*})$ and $(y - x_{m^*})\omega(U_{m^*})$, where x_{m^*} and x_{m^*+1} are the closest members to the observation, i.e. $x_{m^*} \leqslant y \leqslant x_{m^*+1}$.

We illustrate our example for $M = 4$ members, in Figure 4.5. We consider that $x_1 < x_2 < x_3 < x_4$, without loss of generality. For this example, the function ω is convex. Higher importance is given to the levels of quantile close to 0 or 1 than those close to 0.5. The step function $\omega(\mathrm{G}^{\mathcal{E}})$ gives a variable importance to the 3 intervals $[x_1, x_2]$, $[x_2, x_3]$ and $[x_3, x_4]$. For uniformly distributed members $u_1 = u_2 = u_3 = u_4 = 0.25$, the importance $\omega(\mathrm{G}^{\mathcal{E}})$ reaches the values $\omega(0.25)$ on the interval $[x_1, x_2], \omega(0.5)$ on the interval $[x_2, x_3]$ and $\omega(0.75)$ on the interval $[x_3, x_4]$. In our case where ω is symmetrical function $(\omega(\alpha) = \omega(1-\alpha))$, the interval $[x_2, x_3]$ has the lowest importance and the importance of the intervals $[x_1, x_2]$ and $[x_3, x_4]$ are equal. We also show the case of non-uniform weights $u_1 = 0.3$, $u_2 = 0.3$, $u_3 = 0.3$, $u_4 = 0.1$, where the interval $[x_1, x_2]$ has a lower importance than the interval $[x_3, x_4]$. Compared to the example of Section 4.1.1 with threshold-weighted score, the quantile-weighted score also gives a variable importance to domains of $\mathbb{R}$. While the domains of variable importance are

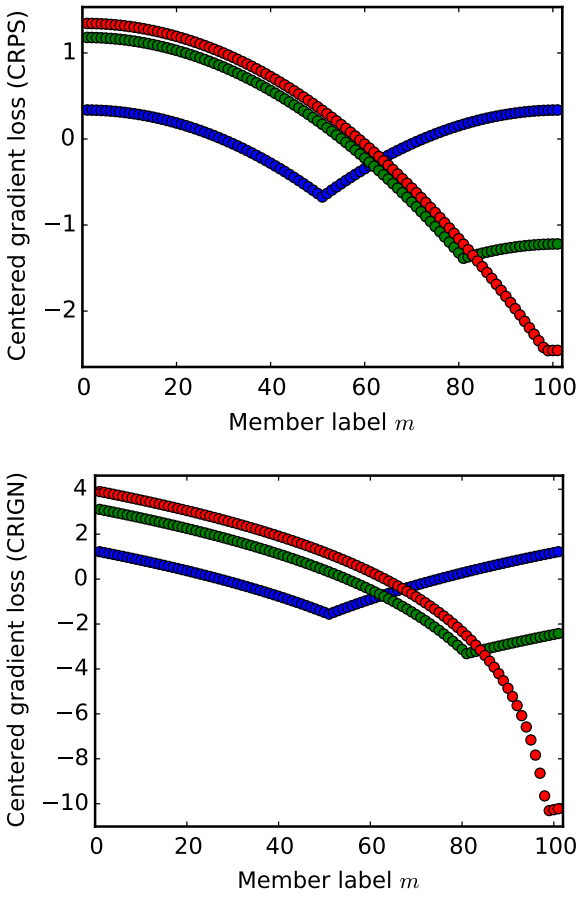


Figure 4.6 – Centered loss gradients $\widetilde{\ell}_m - \sum_{k=1}^{M} u_k \widetilde{\ell}_k$ of the CRPS (top) and the CRIGN (bottom) for M=101 linearly spaced members in [-2, 2] with observations $y = 0$ (blue), $y = 1.2$ (green) and $y = 1.9$ (red).

fixed for threshold-weighted score, the domains of variable importance rely on both u_m and x_m for quantile-weighted scores.

Illustrations of the loss gradients are proposed in Figure 4.6. We picked $M = 101$ linearly spaced members between -2 and 2 and show the loss gradients of the CRPS and the CRIGN for 3 observations $y \in \{0, 1.2, 1.9\}$. We see that the extreme levels of quantile receive relatively lower losses with the CRPS than with the CRIGN due to the higher concavity of the CRPS gradient loss (see the case $y = 0$). Besides, when the observation reaches a distribution tail (see the case $y = 1.9$), the relative magnitude of the CRIGN gradients is higher than the relative magnitude of the CRPS gradients. In other words, the weights u_m learned by CRIGN minimization are more influenced by observations reaching the observations tails.

The study of the bias of quantile-weighted scores is left for future research. Using the notation $\mathrm{G} = \mathrm{E}(\mathrm{G}^{\mathcal{E}})$, the difficulty of the study of the bias in the general case is the comparison of $\mathrm{E}(\beta_1(\mathrm{G}^{\mathcal{E}}))$ and $\mathrm{E}(\beta_0(1 - \mathrm{G}^{\mathcal{E}}))$ to $\beta_1(\mathrm{G})$ and $\beta_0(1 - \mathrm{G})$.

Convexity of quantile-weighted scores

In order to use quantile-weighted scoring rules in sequential aggregation, we study the convexity of the score

$$S(\mathrm{G}, y) = \int \mathrm{H}_y \beta_1(\mathrm{G}) + (1 - \mathrm{H}_y)\beta_0(1 - \mathrm{G})\,,$$

with respect to the weight vector $\boldsymbol{u}$ defining $\mathrm{G} = \sum_{m \leqslant M} u_m \mathrm{G}_m$. We recall that convex loss functions are necessary to get regret bounds against the best fixed combination of experts. The convexity cannot be obtained by claiming a sum of convex functions, because the sum depends on the variable $\boldsymbol{u}$ which is concerned by the convexity.

From the mth gradient $\int \mathrm{G}_m(\mathrm{G} - \mathrm{H}_y)\omega(\mathrm{G})$, we deduce the (m, k) component of the Hessian matrix $\mathbf{A}$

$$A_{m,k} = \frac{\partial^2 S(\mathrm{G}, y)}{\partial u_m \partial u_k} = \int \mathrm{G}_m \mathrm{G}_k(\omega(\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)\omega'(\mathrm{G}))\,, \tag{4.15}$$

with ω' being the derivative of ω. The score is convex if the Hessian matrix is positive definite, namely if for any vector $\boldsymbol{w} \in \mathbb{R}^M$, $\boldsymbol{w}^\top \mathbf{A} \boldsymbol{w} \geqslant 0$.

The condition $\omega(\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)\omega'(\mathrm{G}) \geqslant 0$ is sufficient to prove the convexity since we have

$$\begin{aligned}
\boldsymbol{w}^\top \mathbf{A} \boldsymbol{w} &= \sum_{m,k=1}^{M} w_m w_k A_{m,k} \\
&= \int \left(\sum_{m,k=1}^{M} w_m w_k \mathrm{G}_m \mathrm{G}_k \right) (\omega(\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)\omega'(\mathrm{G})) \\
&= \int \left(\sum_{m=1}^{M} w_m \mathrm{G}_m \right)^2 (\omega(\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)\omega'(\mathrm{G}))\,.
\end{aligned}$$

We find the convexity of the CRPS in the case $\omega = 2$. Besides, we prove the convexity of the CRIGN, obtained in the case $\omega(\alpha) = 1/(\alpha(1-\alpha)) = 1/\alpha + 1/(1-\alpha)$. We recall the CRIGN definition:

$$\mathrm{CRIGN} = -\int \mathrm{H}_y \ln \mathrm{G} + (1 - \mathrm{H}_y) \ln(1 - \mathrm{G})\,.$$

We have

$$\omega'(\alpha) = \frac{-1}{\alpha^2} + \frac{1}{(1-\alpha)^2} = \frac{2\alpha - 1}{\alpha^2(1-\alpha)^2}\,,$$

and

$$\begin{aligned}
\omega(\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)\omega'(\mathrm{G}) &= \frac{1}{\mathrm{G}(1-\mathrm{G})} + (\mathrm{G} - \mathrm{H}_y)\frac{2\mathrm{G} - 1}{\mathrm{G}^2(1-\mathrm{G})^2} \\
&= \frac{1}{\mathrm{G}^2(1-\mathrm{G})^2}(\mathrm{G}(1-\mathrm{G}) + (\mathrm{G} - \mathrm{H}_y)(2\mathrm{G} - 1)) \\
&= \frac{1}{\mathrm{G}^2(1-\mathrm{G})^2}(\mathrm{H}_y - \mathrm{G})^2\,, \\
&\geqslant 0\,,
\end{aligned}$$

which concludes the proof. The last equality holds since H_y is either equal to 0 or 1.

The intermediate situation of $\omega(\alpha) = 1/\sqrt{\alpha(1-\alpha)}$ lies between the CRPS and the CRIGN. The related score is defined by

$$\int \mathrm{H}_y \arcsin(\sqrt{1-\mathrm{G}}) + (1-\mathrm{H}_y)\arcsin(\sqrt{\mathrm{G}}) - \sqrt{\mathrm{G}(1-\mathrm{G})}\,. \tag{4.16}$$

We also show convexity in this situation. Indeed we have

$$\omega'(\alpha) = \frac{-(1-2\alpha)(\alpha(1-\alpha))^{-3/2}}{2}\,,$$

from what we deduce that

$$\begin{aligned}
\omega(\mathrm{G}) + (\mathrm{G}-\mathrm{H}_y)\omega'(\mathrm{G}) &= (\mathrm{G}(1-\mathrm{G}))^{-3/2}(\mathrm{G}(1-\mathrm{G}) - \frac{1}{2}(\mathrm{G}-\mathrm{H}_y)(1-2\mathrm{G}) \\
&= (\mathrm{G}(1-\mathrm{G}))^{-3/2}(\mathrm{H}_y(1-\mathrm{G})/2 + (1-\mathrm{H}_y)\mathrm{G}/2) \\
&\geqslant 0\,,
\end{aligned}$$

using the same trick that H_y is either equal to 0 or 1.

Conclusion

Threshold-weighted and quantile-weighted scores allow to focus on specific areas of interest on the real line or in the forecaster's distribution. We showed that the results of Thorey et al. [TMB16] can be generalized to threshold-weighted scoring rules. Indeed, the effect of the threshold-weighting is equivalent to a simple data transformation. The bias of the ensemble CRPS and the definition of the class CRPS therefore apply to the transformed data. Besides, we demonstrated several results for quantile-weighted scoring rules. After rewriting the score definition, the optimal locations for an ensemble of forecasts were derived. The loss gradients of a quantile-weighted score can be interpreted with a data transformation, but contrarily to the threshold-weighted scoring rules, the data transformation depends on the forecaster's distribution. However, the study of the bias of quantile-weighted scoring rules is left for future research.

Other areas of further work include the study of quantile-weighted scoring rules for extreme value verification, with a focus on the distribution tails. The CRIGN may not be a viable option because this score is not bounded even with bounded observational distribution. A similar study would of interest based on expectiles instead of quantiles. Besides, analytic expressions of the scores for parametric distributions and mixture of parametric distributions would facilitate the testing of these new scores. Evaluation and statistical learning with quantile-weighted scoring rules for real world data sets are shown in Appendix 7.A.

4.2 Probabilistic forecasting with observational noise

Bibliographical remarks

We investigate how a forecaster can take into account some knowledge about noise or perturbed observations in the setting of ensemble forecasting. A major issue in this field of investigation is that the noise distribution is generally unknown. The topic of noisy observations is widely studied in the data assimilation community, where a central question is to provide the best trade-off between new observations and background knowledge, both imperfect. When the observation errors are uncorrelated, the observation errors are often scaled by the standard deviation of the noise, such as for the Kalman filter. Besides, taking into account the noise by either perturbing the observation or perturbing the members at the corresponding level of noise can improve the forecasts verification [Sae+04; CT08]. In the case of independent additive noise on the observations, Bowler [Bow06] and Bowler [Bow08] demonstrate deconvolution methods to denoise the received observations. They rely on the fact that the distribution of a sum of independent random variables is the convolution of their PDFs.

We emphasize that the setting of noisy observations is seldom investigated in the framework of online learning with experts advice. The losses of each expert are usually considered i.i.d. or noiseless, but rarely with time-varying noise distributions. Still, Yang [Yan04] shows regret bounds on the loss against the observation expectations (and not against the noisy observations) for the square loss. Interestingly, a normalization of the loss by the observation standard deviation is proposed, as in penalized least square regression. Cesa-Bianchi et al. [CSS11] focus mainly on expert noise and obtain unbiased estimates of the loss gradients for regret bounds in expectation. We discuss below the interest of the loss expectation in the context of probabilistic forecasting with noisy observations. In online classification with label noise, the work of Ben-David et al. [BPS09] shows regret bounds in expectation by considering the probability that the label shown to the forecaster is correct or not. Similar approaches are also studied in the batch setting [Nat+13], see Frénay and Verleysen [FV14] for a review of classification with label noise.

Context

Say the forecaster receives a distribution or multiple observations instead of a single observation. A basic idea would be to switch from the CRPS to the average CRPS, where the expectation is taken according to the distribution of the observations. Let G be the forecasted CDF, F be the CDF of the observations received by the forecaster and Y be a random variable described by F. The average CRPS is equal to the squared difference between F and G, plus the uncertainty term of F:

$$\mathrm{E}(\mathrm{CRPS}(\mathrm{G}, \mathrm{Y})) = \int (\mathrm{F} - \mathrm{G})^2 + \int \mathrm{F}(1 - \mathrm{F})\,.$$

Two elements should be considered:

(i) the reality, that we note y^t, is not a random variable, but a fixed value. The superscript t here stands for "true". We emphasize that the distribution G should

reflect the inability of the forecaster to provide perfect estimations of y^t, whatever the level of observational noise. The assumed distribution of y^t is in fact the most informative distribution that a forecaster could deliver.

(ii) the observation y may be corrupted by some noise, from an inaccurate sensor for instance. If this noise is very large, the forecaster may not be willing to depict such a large uncertainty in its forecasts. In any case, the forecaster wishes to predict the reality y^t and not the observation y.

Consequently, the forecaster may look for a different objective than targeting the observational CDF F to compute the weights of the model mixture $\mathrm{G} = \sum_{m \leqslant M} u_m \mathrm{G}_m$. We now give thought on strategies for the forecaster wishing both to apply ensemble post-processing methods and evaluate the resulting predictions. We include the aforementioned strategies in our list:

— *Empirical normalization*: scale the members and the observations by an empirical value reflecting the level of noise, such as the standard deviation of the noise. Less weight is attributed to time steps with large inaccuracy. Besides Yang [Yan04], we did not find theoretical guarantee and online learning algorithms taking these normalizing factors into account. For probabilistic forecasting, we suggest to normalize the data by the uncertainty term of the loss function, which is the Gini mean difference of the observations for the CRPS.

— *Forecasting noisy observations*: playing the distribution of the noisy observations F, with the loss $\int (\mathrm{F} - \mathrm{G})^2$. This strategy is of limited interest because of the above point (ii).

— *Perturbing ensemble members*: generate perturbed ensemble members according to the observational noise, and work with the perturbed members. However, if the perturbed distribution of the members match the observational distribution, it is not ensured that the distribution of the unperturbed members match the underlying distribution of the truth, as stated by Bröcker and Smith [BS07b] in terms of scoring rules.

— *Noise deconvolution*: generate denoised observations from deconvolution of the observational distribution.

— *Forecasting reality and not the observations*: try to forecast samples from F, and not directly F. The motivation behind is that the reality y^t is most likely located in domains with high observational density. We introduce a method to do so in Section 4.2.1.

4.2.1 Generalized least square with the CRPS

Let M forecasts be described by the CDFs G_m, and $\mathcal{Y}_{\mathrm{CDF}}$ be the space of the CDFs. In common data assimilation formulations, the observation state contains several scalar observations, while our observation state describes the occurrence of threshold exceeding of only one observation. An observation y is described by the unit step function $\mathrm{H}_y \in \mathcal{Y}_{\mathrm{CDF}}$ centered on y. The CRPS between $\mathrm{G} = \sum_{m \leqslant M} u_m \mathrm{G}_m$ and y defined by $\int (\mathrm{H}_y - \mathrm{G})^2$ can be written with the inner product on functions:

$$(\mathrm{H}_y - \mathrm{G})^\top (\mathrm{H}_y - \mathrm{G}) = \int (\mathrm{H}_y - \mathrm{G})^2 \,. \tag{4.17}$$

We wish to obtain the best estimate of $\boldsymbol{u}^*$ defined by the model

$$\mathrm{H}_y = \sum_{m \leqslant M} u_m^* \mathrm{G}_m + \boldsymbol{e}^o \,, \tag{4.18}$$

where $\boldsymbol{e}^o$ is the observational error. We emphasize that in this model, the first objective of the forecaster is to be close to the CDF of a unit step function. In this sense, the forecaster targets samples from the distribution described by F and not directly F. The assumption of model unbiasedness forces the equality $\sum_{m \leqslant M} u_m^* \mathrm{G}_m = \mathrm{F}$ because $\mathrm{E}(\mathrm{H}_y) = \mathrm{F}$. Consequently, $\boldsymbol{e}^o = \mathrm{H}_y - \mathrm{F}$, and the covariance matrix of the observational errors $\mathrm{R} = \mathrm{E}\left[\boldsymbol{e}^o (\boldsymbol{e}^o)^\top\right]$ is fully determined by F, whose knowledge is required.

We now choose a rule to find a good estimate of $\boldsymbol{u}^*$, in the form of a minimization solution. Let R^{-1} be the matrix inverse of R, then a good estimate of $\boldsymbol{u}$ is given by the minimizer of the generalized least squares $(\mathrm{H}_y - \mathrm{G})^\top \mathrm{R}^{-1} (\mathrm{H}_y - \mathrm{G})$, as derived by [Ait36]. The idea is to shift the minimization problem in a space with uncorrelated, homoscedastic errors, and apply ordinary least square (OLS) in this space. Therefore, the theoretical guarantees of OLS apply in the transformed space. For example, the OLS estimator is unbiased and efficient. The solution of the OLS minimization has the minimum variance compared to other unbiased linear estimators. In data assimilation, the matrix R is commonly assumed to be diagonal, i.e. the errors are uncorrelated. This assumption does not hold in our case, see below the formula of R.

As we show below, this new method allows to target the distribution of the observations, while being quite different from minimizing the average CRPS. We relate the score expectation to the value of Pearson's χ^2 test statistic. We inform the reader that many open questions remain at the end of this chapter, which is present in this thesis because our preliminary theoretical results look promising.

Observational error covariance matrix R

Now we derive formulas for R and R^{-1}. In a continuous observation space, R is a symmetrical covariance operator such that

$$\begin{aligned}
\mathrm{R}(x,z) &= \mathrm{E}_\mathrm{Y}\left[(\mathrm{H}_\mathrm{Y}(x) - \mathrm{F}(x))(\mathrm{H}_\mathrm{Y}(z) - \mathrm{F}(z))\right] \\
&= \mathrm{E}_\mathrm{Y}\left[\mathrm{H}_\mathrm{Y}(x)\mathrm{H}_\mathrm{Y}(z)\right] - \mathrm{F}(x)\mathrm{F}(z) \\
&= \mathrm{E}_\mathrm{Y}\left[\mathrm{H}_\mathrm{Y}(\min(x,z))\right] - \mathrm{F}(x)\mathrm{F}(z) \\
&= \mathrm{F}(\min(x,z)) - \mathrm{F}(x)\mathrm{F}(z) \\
&= \mathrm{F}(\min(x,z))(1 - \mathrm{F}(\max(x,z)))\,,
\end{aligned}$$

using $\mathrm{E}_\mathrm{Y}(\mathrm{H}_\mathrm{Y}) = \mathrm{F}$. We note that the quantity $\mathrm{R}(x,z)$ reaches a maximum for $\mathrm{F}(x) = \mathrm{F}(z) = 0.5$ and that the diagonal terms $\mathrm{R}(x,x)$ are higher than the off-diagonal terms $\mathrm{R}(x,z)$ for any x and $z \neq x$.

Interestingly, the covariance operator is related to the Gini mean difference and to the variance of the random variable Y described by F:

$$\mathrm{Tr}(\mathrm{R}) = \int \mathrm{R}(x,x)\mathrm{d}x = \int \mathrm{F} - \mathrm{F}^2 = \frac{1}{2}\,\mathrm{E}_{\mathrm{Y},\mathrm{Y}'}(|\mathrm{Y} - \mathrm{Y}'|)\,, \tag{4.19}$$

and

$$\iint \mathrm{R}(x,z)\mathrm{d}x\mathrm{d}z = \frac{1}{2}\,\mathrm{E}_{\mathrm{Y},\mathrm{Y}'}((\mathrm{Y}-\mathrm{Y}')^2) = \mathrm{E}_{\mathrm{Y}\sim\mathrm{F}}((\mathrm{Y}-\bar{\mathrm{Y}})^2)\,, \tag{4.20}$$

where $\bar{\mathrm{Y}} = \mathrm{E}(\mathrm{Y})$. Equation 4.20 was found in González Abril et al. [Gon+10].

For the sake of simplicity, we now consider a grid $z_1 < z_2 < ... < z_S$ discretizing the real line, with $0 < \mathrm{F}(z_1) < \cdots < \mathrm{F}(z_j) < \cdots < \mathrm{F}(z_S) < 1$. Working in this simplified discrete case where $\mathcal{Y}_{\mathrm{CDF}}$ has a finite dimension allows us to derive expressions for both R and R^{-1}. A natural choice is to pick regularly spaced quantiles verifying $\mathrm{F}(z_{j+1}) - \mathrm{F}(z_j) = 1/(S+1)$. We describe the symmetrical matrix R by its components:

$$R_{i,j} = \min(\mathrm{F}_i, \mathrm{F}_j)(1 - \max(\mathrm{F}_i, \mathrm{F}_j))\,, \tag{4.21}$$

with the notation $\mathrm{F}_j = \mathrm{F}(z_j)$. The said inverse R^{-1} is tridiagonal with the following components on the ith line:

$$\left[0\dots0 \quad \underbrace{\frac{-1}{F_i - F_{i-1}}}_{i-1} \quad \underbrace{\frac{1}{F_i - F_{i-1}} + \frac{1}{F_{i+1} - F_i}}_{i} \quad \underbrace{\frac{-1}{F_{i+1} - F_i}}_{i+1} \quad 0\dots0\right]$$

with a slight difference for the first and the last line, respectively:

$$\left[\frac{\mathrm{F}_2}{\mathrm{F}_1(\mathrm{F}_2 - \mathrm{F}_1)} \quad \frac{-1}{\mathrm{F}_2 - \mathrm{F}_1} \quad 0\dots0\right],$$

and

$$\left[0\dots0 \quad \frac{-1}{\mathrm{F}_S - \mathrm{F}_{S-1}} \quad \frac{1 - \mathrm{F}_{S-1}}{(1 - \mathrm{F}_S)(\mathrm{F}_S - \mathrm{F}_{S-1})}\right].$$

The general expression of the diagonal term is valid for all indices i with the notation, $\mathrm{F}_0 = 0$, $\mathrm{F}_{S+1} = 1$, $z_0 = -\infty$, and $z_{S+1} = +\infty$. Appendix 4.A exhibits a derivation of the general term of R^{-1}. The term $F_{j+1} - F_j$ is the probability of the event "$\mathrm{Y} \in B_j$", where B_j is the bin $[z_j, z_{j+1}]$. The presence of such terms in R^{-1} suggests that our penalized least square procedure is related to the occurrence of the events "$\mathrm{Y} \in B_j$".

Examples of R and R^{-1} are shown in Figure 4.7 in the cases of uniform, piecewise uniform and Gaussian distributions for linearly spaced thresholds. We see that R is determined by F, while R^{-1} is determined by the slope of F. Steep gradients of F are related to low values in R^{-1}, while moderate gradients are related to high values in R^{-1}. The example of the Gaussian distribution shows the huge impact of very low gradients in the tails of the distribution. The case of the uniform distribution also illustrates the situation of linearly spaced quantiles (instead of thresholds), where equal probability is given to each bin ($\mathrm{F}(z_{j+1}) - \mathrm{F}(z_j) = 1/(S+1)$).

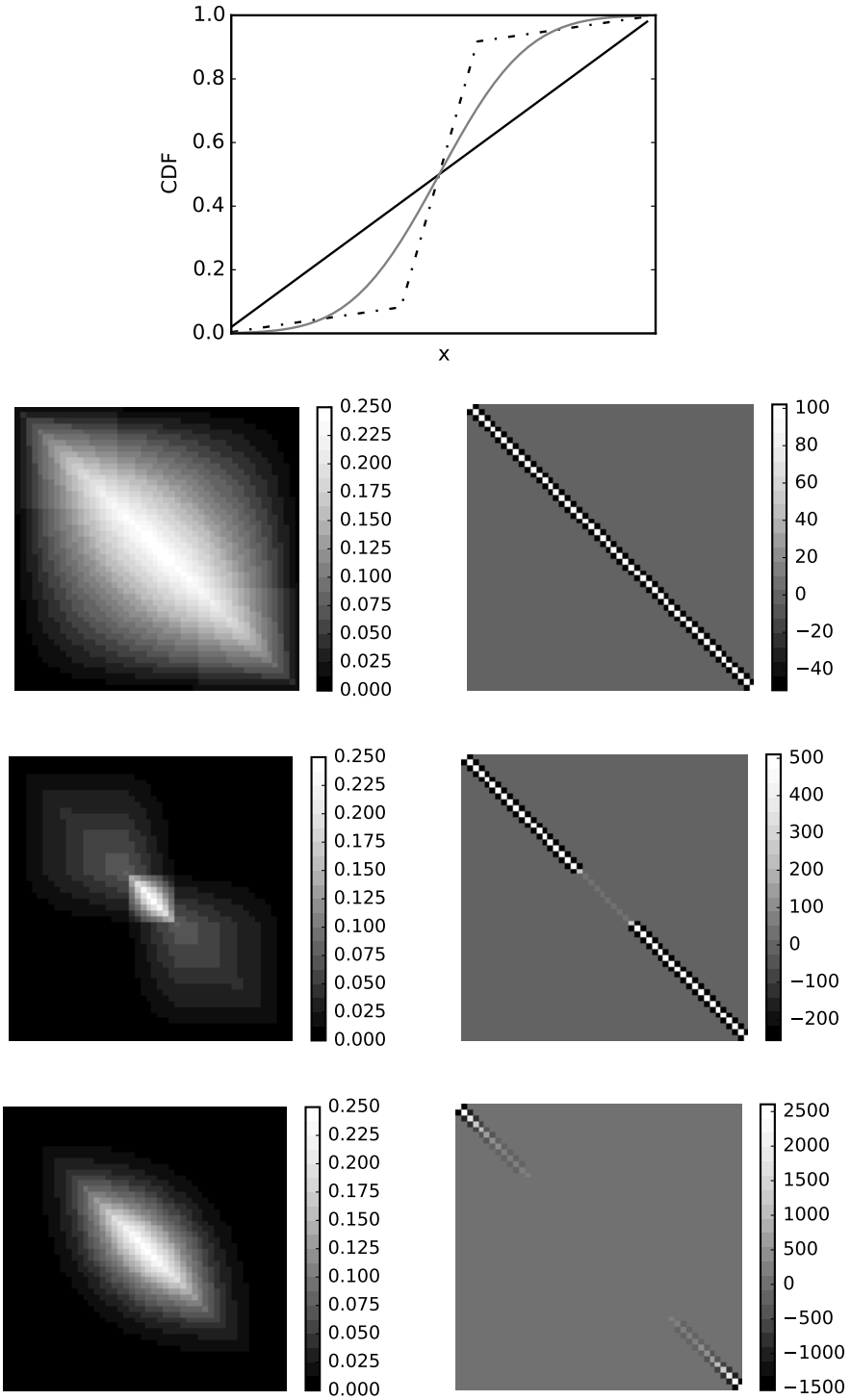


Figure 4.7 – Visualization of R (left) and R^{-1} (right) in the cases of uniform (top), piecewise uniform (middle) and Gaussian distribution (bottom) for linearly spaced thresholds. The CDFs of the uniform (solid black), piecewise uniform (dashed black) and Gaussian distribution (solid grey) are also shown.

A new loss related in expectation to the χ^2 test

Now we compute the loss $\ell^{\mathrm{R}^{-1}}(\boldsymbol{u}) = (\boldsymbol{e})^\top \mathrm{R}^{-1}(\boldsymbol{e})$, where $\boldsymbol{e}$ is a difference of discretized CDFs verifying $\boldsymbol{e}_0 = \boldsymbol{e}_{S+1} = 0$. We have

$$\begin{aligned}
\ell^{\mathrm{R}^{-1}}(\boldsymbol{u}) &= \sum_{j=1}^{S} \boldsymbol{e}_j \left[\frac{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}}{\mathrm{F}_{j+1} - \mathrm{F}_j} + \frac{\boldsymbol{e}_j - \boldsymbol{e}_{j-1}}{\mathrm{F}_j - \mathrm{F}_{j-1}} \right] \\
&= \sum_{j=1}^{S} \boldsymbol{e}_j \frac{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}}{\mathrm{F}_{j+1} - \mathrm{F}_j} + \sum_{j=1}^{S} \boldsymbol{e}_{j+1} \frac{\boldsymbol{e}_{j+1} - \boldsymbol{e}_j}{\mathrm{F}_{j+1} - \mathrm{F}_j} + \frac{\boldsymbol{e}_1^2}{\mathrm{F}_1} \\
&= \sum_{j=1}^{S} \frac{(\boldsymbol{e}_{j+1} - \boldsymbol{e}_j)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j} + \frac{\boldsymbol{e}_1^2}{\mathrm{F}_1} \\
&= \sum_{j=0}^{S} \frac{(\boldsymbol{e}_{j+1} - \boldsymbol{e}_j)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j} \,.
\end{aligned}$$

In the last line, the term $j = 0$ is $\frac{\boldsymbol{e}_1^2}{\mathrm{F}_1}$ and the term $j = S$ is $\frac{\boldsymbol{e}_S^2}{(1-\mathrm{F}_S)}$. We emphasize that this expression is valid for any $\boldsymbol{e}$ verifying $\boldsymbol{e}_0 = \boldsymbol{e}_{S+1} = 0$, hence the errors $\boldsymbol{e}_i$ may be defined by either $\boldsymbol{e}_i = \mathrm{H}_y(z_i) - \mathrm{G}(z_i)$ or $\boldsymbol{e}_i = \mathrm{F}(z_i) - \mathrm{G}(z_i)$. When F describes the climatological distribution and an observation y is received, the forecaster may be more willing to compare G with H_y than with F.

At first look, the new loss could seem to be closely related to the CRPS, up to normalization factors $\mathrm{F}_{j+1} - \mathrm{F}_j$. A normalized Ranked Probability Score (RPS) may be obtained by taking only the diagonal terms of R^{-1}, giving $\sum_{j=1}^{S} \frac{(\boldsymbol{e}_j)^2}{\mathrm{F}_j(1-\mathrm{F}_j)}$. However, the loss $\ell^{\mathrm{R}^{-1}}(\boldsymbol{u})$ is minimized when the difference $\boldsymbol{e}_{j+1} - \boldsymbol{e}_j$ of errors is equal to the difference $\mathrm{F}_{j+1} - \mathrm{F}_j$ for each j. We explain below why this is a major difference with the CRPS.

We restrict our study to the case of ensemble forecasting, where the CDF $\mathrm{G}^{\mathcal{E}}$ is a sum of unit step functions centered on the members x_m. The difference of errors $\boldsymbol{e}_{j+1} - \boldsymbol{e}_j$ is then related to the presence of x_m and y in the bin B_j. Let $\mathbf{1}_{B_j}$ be the indicator function of B_j ($\mathbf{1}_{B_j}(z) = 1$ if $z \in B_j$ and zero otherwise), and let $u^{(j)} = \mathrm{G}^{\mathcal{E}}(z_{j+1}) - \mathrm{G}^{\mathcal{E}}(z_j) = \sum_{m \leqslant M} \mathbf{1}_{B_j}(x_m) u_m$ be the cumulated weights of the members in B_j. For example, if x_4, x_7 and y are in the bin B_j and no other members are in this bin, then $\boldsymbol{e}_{j+1} - \boldsymbol{e}_j = 1 - u_4 - u_7$. If x_2 is the only member in B_j and the observation is not in B_j, then $\boldsymbol{e}_{j+1} - \boldsymbol{e}_j = -u_2$.

The loss may be rewritten as

$$\ell^{\mathrm{R}^{-1}}(\boldsymbol{u}) = \sum_{j=0}^{S} \frac{\left(\mathbf{1}_{B_j}(y) - u^{(j)}\right)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j} = \sum_{j=0}^{S} \frac{\mathbf{1}_{B_j}(y) \left(1 - u^{(j)}\right)^2 + (1 - \mathbf{1}_{B_j}(y)) \left(u^{(j)}\right)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j} \tag{4.22}$$

This score can be seen as a discretized quadratic score * (taking only the numerators) with normalizing factors.

∗. The quadratic score of the PDF g against the observation y is $-2g(y) + \int g^2$.

According to the location of the members and the location of the observation, the loss $\ell^{\mathrm{R}^{-1}}(\boldsymbol{u})$ simplifies to:

$$\ell^{\mathrm{R}^{-1}}(\boldsymbol{u}) = \frac{\left(1-u^{(j^*)}\right)^2}{\mathrm{F}(z_{j^*+1})-\mathrm{F}(z_{j^*})} + \sum_{\substack{j=0\\ j\neq j^*}}^{S} \frac{\left(u^{(j)}\right)^2}{\mathrm{F}(z_{j+1})-\mathrm{F}(z_j)}\,, \tag{4.23}$$

where the bin B_{j^*} contains the observation. In terms of weight optimization, two different situations may occur:

— Case 1: the bin B_{j^*} contains the observation and at least one member. The optimal weights verify $u^{(j^*)} = 1$ and $u^{(j\neq j^*)} = 0$. In other words, the members outside of the correct bin should receive a null weight.

— Case 2: the bin B_{j^*} contains the observation, but does not contain any member ($u^{(j^*)} = 0$). The loss is minimized when $u^{(j)} = \mathrm{F}(z_{j+1}) - \mathrm{F}(z_j)$ for all j such that the bin B_j contains at least one member. Interestingly, the optimal weights do not depend on the location of the observation.

To summarize the above cases, the members should try to be in the correct bin, but if no member is good enough to be in the correct bin, then the weights should be in agreement with F. We emphasize that the penalization of the members lying outside of the correct bin is not sensitive to the distance to the observation, which is in sharp contrast with the CRPS. Besides, the normalizing factors scale the loss related to $u^{(j)}$ with the probability that the observation is in B_j. Interestingly, the tridiagonal structure $[-1, 2, -1]$ of R^{-1} recalls two differentiations, one for each $\boldsymbol{e}$ in $\ell^{\mathrm{R}^{-1}}(\boldsymbol{u})$ one may say. We see that the generalized least square procedure shifts the problem from the CDF point of view to the PDF point of view.

We now check that the score is proper. Strict propriety is not obtained because of the discretization, but we show that the score is minimized by a CDF G verifying $\mathrm{G}(z_j) = \mathrm{F}(z_j)$. Indeed, by taking the expectation over the observational distribution:

$$\begin{aligned}
\mathrm{E}\left[(\boldsymbol{e})^\top \mathrm{R}^{-1}(\boldsymbol{e})\right] &= \sum_{j=0}^{S} \frac{(\mathrm{F}_{j+1}-\mathrm{F}_j)\left(1-u^{(j)}\right)^2 + (1-(\mathrm{F}_{j+1}-\mathrm{F}_j))\left(u^{(j)}\right)^2}{\mathrm{F}_{j+1}-\mathrm{F}_j} \\
&= \sum_{j=0}^{S} \frac{\left(\mathrm{F}_{j+1}-\mathrm{F}_j-u^{(j)}\right)^2}{\mathrm{F}_{j+1}-\mathrm{F}_j} + \frac{(\mathrm{F}_{j+1}-\mathrm{F}_j)\,(1-(\mathrm{F}_{j+1}-\mathrm{F}_j))}{\mathrm{F}_{j+1}-\mathrm{F}_j} \\
&= \sum_{j=0}^{S} \frac{\left(\mathrm{F}_{j+1}-\mathrm{F}_j-u^{(j)}\right)^2}{\mathrm{F}_{j+1}-\mathrm{F}_j} + S\,.
\end{aligned}$$

The proof is complete since the optimal distribution G verifies $\mathrm{G}(z_{j+1}) - \mathrm{G}(z_j) = \mathrm{F}(z_{j+1}) - \mathrm{F}(z_j)$ and $\mathrm{G}(z_0) = \mathrm{F}(z_0) = 0$. The uncertainty term of the score is equal to S, which depends only on the grid resolution, and not on the observational distribution. Besides, the divergence term of $\mathrm{E}\left[(\boldsymbol{e})^\top \mathrm{R}^{-1}(\boldsymbol{e})\right]$ is equal to Pearson's χ^2 test-statistic. Connections between χ^2 statistics and generalized least squares are not new [Rao02]. To the best of our knowledge, the approach of generalized least squares applied to discretized CDFs is innovative. The question of score bias is addressed in Appendix 4.A.

4.2.2 Discussion and further work

Assessing the flatness of the rank histogram is possible with the χ^2 statistic [And96], even though such verification tool may not be adequate because it treats all bins equally and forgets the rank structure [Elm05]. In any case, minimizing the new loss $\ell^{\mathrm{R}^{-1}}(\boldsymbol{u})$ may be seen as a rank histogram optimization procedure.

The χ^2 statistic is quite celebrated. Providing new insights on this statistic therefore seemed of interest to us. Further work could investigate connections with Cramer-von Mises criterion $\int(\mathrm{G}-\mathrm{F})^2\mathrm{dF}$ and Anderson-Darling statistic $\int \frac{(\mathrm{G}-\mathrm{F})^2\mathrm{dF}}{\mathrm{F}(1-\mathrm{F})}$. Indeed, they appear to be related to simplifications of the χ^2 statistic by not taking R^{-1} into account or only partially with the diagonal only.

The distribution of the observations F is in general unknown. We emphasize that the lack of knowledge of F does not change the score non-strict propriety. Indeed, if the loss is built with the CDF F^{guess} and the observations follow the CDF F, the scoring rule is still minimized in average if the forecaster's CDF follows F and not F^{guess}. Using an incorrect F^{guess} may not impair the scoring rule.

It would be interesting to test this new loss, for both learning (in online learning for example) and evaluating probabilistic forecasts. For real world data sets, the first task would be to determine the observational distribution. The climatological distribution of the observations could be a good starting point. Secondly, the forecaster should determine the appropriate thresholds z_j or equivalently the bins B_j. A rule of thumb for Pearson's test states that each bin should contain at least 5 observations. Photovoltaic power data sets may not be a good starting point, because observational uncertainty is seldom taken into account contrarily to other variables such as geopotential heights or temperature [Sae+04; CT08]. Further work may investigate the integration of such climatological knowledge in the loss function, and up to which level the forecaster may benefit from the integration of this knowledge.

We conjecture that theoretical results may possibly be obtained by using this new loss in sequential aggregation, for example showing a reduced variance of the weights due to the knowledge of observation uncertainty.

Appendix 4.A Supplementary material

Proof of R inverse formula in the discretized case:

Here we demonstrate that the matrix noted R^{-1} is indeed the inverse of R. We make an intensive use of $\mathrm{R}_{i,j} = \mathrm{F}_{\min(i,j)}(1-\mathrm{F}_{\max(i,j)})$ and of the matrix product formula $(\mathrm{AB})_{i,j} = \sum_{k\leqslant S} \mathrm{A}_{i,k}\mathrm{B}_{k,j}$.

First we compute $(\mathrm{R}\mathrm{R}^{-1})_{i,j}$ for $1 < j < i \leqslant S$,

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{i,j} &= \mathrm{R}_{i,j-1}\mathrm{R}^{-1}_{j-1,j} + \mathrm{R}_{i,j}\mathrm{R}^{-1}_{j,j} + \mathrm{R}_{i,j+1}\mathrm{R}^{-1}_{j+1,j} \\ &= \frac{-\mathrm{F}_{j-1}(1-\mathrm{F}_i)}{\mathrm{F}_j - \mathrm{F}_{j-1}} + \frac{\mathrm{F}_j(1-\mathrm{F}_i)(\mathrm{F}_{j+1}-\mathrm{F}_{j-1})}{(\mathrm{F}_{j+1}-\mathrm{F}_j)(\mathrm{F}_j-\mathrm{F}_{j-1})} - \frac{\mathrm{F}_{j+1}(1-\mathrm{F}_i)}{\mathrm{F}_{j+1}-\mathrm{F}_j} \\ &= \frac{1-\mathrm{F}_i}{(\mathrm{F}_{j+1}-\mathrm{F}_j)(\mathrm{F}_j-\mathrm{F}_{j-1})} \\ &\quad \times \left(-\mathrm{F}_{j-1}(\mathrm{F}_{j+1}-\mathrm{F}_j) + \mathrm{F}_j(\mathrm{F}_{j+1}-\mathrm{F}_{j-1}) - \mathrm{F}_{j+1}(\mathrm{F}_j-\mathrm{F}_{j-1})\right) \\ &= 0\,,\end{aligned}$$

and for $1 \leqslant i < j < S$,

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{i,j} &= \mathrm{R}_{i,j-1}\mathrm{R}^{-1}_{j-1,j} + \mathrm{R}_{i,j}\mathrm{R}^{-1}_{j,j} + \mathrm{R}_{i,j+1}\mathrm{R}^{-1}_{j+1,j} \\ &= \frac{-\mathrm{F}_i(1-\mathrm{F}_{j-1})}{\mathrm{F}_j - \mathrm{F}_{j-1}} + \frac{\mathrm{F}_i(1-\mathrm{F}_j)(\mathrm{F}_{j+1}-\mathrm{F}_{j-1})}{(\mathrm{F}_{j+1}-\mathrm{F}_j)(\mathrm{F}_j-\mathrm{F}_{j-1})} - \frac{\mathrm{F}_i(1-\mathrm{F}_{j+1})}{\mathrm{F}_{j+1}-\mathrm{F}_j} \\ &= \frac{\mathrm{F}_i}{(\mathrm{F}_{j+1}-\mathrm{F}_j)(\mathrm{F}_j-\mathrm{F}_{j-1})} \\ &\quad \times \left(-(1-\mathrm{F}_{j-1})(\mathrm{F}_{j+1}-\mathrm{F}_j) + (1-\mathrm{F}_j)(\mathrm{F}_{j+1}-\mathrm{F}_{j-1}) - (1-\mathrm{F}_{j+1})(\mathrm{F}_j-\mathrm{F}_{j-1})\right) \\ &= 0\,.\end{aligned}$$

The above derivations are also valid for $j = 1$ with $\mathrm{F}_0 = 0$ and for $j = S$ with $\mathrm{F}_{S+1} = 1$. Indeed

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{i,1} &= \mathrm{R}_{i,1}\mathrm{R}^{-1}_{1,1} + \mathrm{R}_{i,2}\mathrm{R}^{-1}_{2,1} \\ &= \frac{\mathrm{F}_1(1-\mathrm{F}_i)\mathrm{F}_2}{(\mathrm{F}_2-\mathrm{F}_1)\mathrm{F}_1} - \frac{\mathrm{F}_2(1-\mathrm{F}_i)}{\mathrm{F}_2-\mathrm{F}_1} \\ &= 0\,,\end{aligned}$$

and

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{i,S} &= \mathrm{R}_{i,S-1}\mathrm{R}^{-1}_{S-1,S} + \mathrm{R}_{i,S}\mathrm{R}^{-1}_{S,S} \\ &= \frac{-\mathrm{F}_i(1-\mathrm{F}_{S-1})}{\mathrm{F}_S - \mathrm{F}_{S-1}} + \frac{\mathrm{F}_i(1-\mathrm{F}_S)(1-\mathrm{F}_{S-1})}{(1-\mathrm{F}_S)(\mathrm{F}_S-\mathrm{F}_{S-1})} \\ &= 0\,.\end{aligned}$$

Second we compute $(\mathrm{R}\mathrm{R}^{-1})_{i,i}$ for $1 < i < S$. We have

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{i,i} &= \mathrm{R}_{i,i-1}\mathrm{R}^{-1}_{i-1,i} + \mathrm{R}_{i,i}\mathrm{R}^{-1}_{i,i} + \mathrm{R}_{i,i+1}\mathrm{R}^{-1}_{i+1,i} \\ &= \frac{-\mathrm{F}_{i-1}(1-\mathrm{F}_i)}{\mathrm{F}_i - \mathrm{F}_{i-1}} + \frac{\mathrm{F}_i(1-\mathrm{F}_i)(\mathrm{F}_{i+1}-\mathrm{F}_{i-1})}{(\mathrm{F}_{i+1}-\mathrm{F}_i)(\mathrm{F}_i-\mathrm{F}_{i-1})} - \frac{\mathrm{F}_i(1-\mathrm{F}_{i+1})}{\mathrm{F}_{i+1}-\mathrm{F}_i} \\ &= \frac{1}{(\mathrm{F}_{i+1}-\mathrm{F}_i)(\mathrm{F}_i-\mathrm{F}_{i-1})} \\ &\quad \times \left[-\mathrm{F}_{i-1}(1-\mathrm{F}_i)(\mathrm{F}_{i+1}-\mathrm{F}_i) + \mathrm{F}_i(1-\mathrm{F}_i)(\mathrm{F}_{i+1}-\mathrm{F}_{i-1}) - \mathrm{F}_i(1-\mathrm{F}_{i+1})(\mathrm{F}_i-\mathrm{F}_{i-1})\right] \\ &= 1\,,\end{aligned}$$

using $\mathrm{F}_i(1-\mathrm{F}_{i-1}) - \mathrm{F}_{i-1}(1-\mathrm{F}_i) = \mathrm{F}_i - \mathrm{F}_{i-1}$.

The terms $(\mathrm{R}\mathrm{R}^{-1})_{1,1}$ and $(\mathrm{R}\mathrm{R}^{-1})_{S,S}$ are also equal to one:

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{1,1} &= \mathrm{R}_{1,1}\mathrm{R}^{-1}_{1,1} + \mathrm{R}_{1,2}\mathrm{R}^{-1}_{2,1} \\ &= \frac{\mathrm{F}_1(1-\mathrm{F}_1)\mathrm{F}_2}{(\mathrm{F}_2-\mathrm{F}_1)\mathrm{F}_1} - \frac{\mathrm{F}_1(1-\mathrm{F}_2)}{\mathrm{F}_2-\mathrm{F}_1} \\ &= 1\,,\end{aligned}$$

and

$$\begin{aligned}(\mathrm{R}\mathrm{R}^{-1})_{S,S} &= \mathrm{R}_{S,S-1}\mathrm{R}^{-1}_{S-1,S} + \mathrm{R}_{S,S}\mathrm{R}^{-1}_{S,S} \\ &= \frac{-\mathrm{F}_{S-1}(1-\mathrm{F}_S)}{\mathrm{F}_S-\mathrm{F}_{S-1}} + \frac{\mathrm{F}_S(1-\mathrm{F}_S)(1-\mathrm{F}_{S-1})}{(1-\mathrm{F}_S)(\mathrm{F}_S-\mathrm{F}_{S-1})} \\ &= 1\,,\end{aligned}$$

which completes the proof.

Bias of $(\boldsymbol{e})^\top \mathrm{R}^{-1}(\boldsymbol{e})$.

Let the members X_m be independent random samples described by the CDF G_m, and let $P_m^{(j)}$ be the probability that the mth member is in the bin B_j. The derivations for the ensemble CRPS bias of Section 3.1.4 were achieved for the events $\mathrm{X}_m \leqslant \theta$, while in the current case, the events are of the type $z_j \leqslant \mathrm{X}_m \leqslant z_{j+1}$. We show that similar derivations can be produced to show that $(\boldsymbol{e})^\top \mathrm{R}^{-1}(\boldsymbol{e})$ is a biased score.

We recall Equation 4.22:

$$\ell^{\mathrm{R}^{-1}}(\boldsymbol{u}) = \sum_{j=0}^{S} \frac{\left(\mathbf{1}_{B_j}(y) - u^{(j)}\right)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j}\,.$$

By taking the expectation over the members,

$$\begin{aligned}\mathrm{E}\left[u^{(j)}\right] &= \mathrm{E}\left[\sum_m u_m \mathbf{1}_{B_j}(\mathrm{X}_m)\right] \\ &= \sum_m u_m P_m^{(j)} \\ &= \mathrm{G}(z_{j+1}) - \mathrm{G}(z_j)\,.\end{aligned}$$

Besides,

$$\begin{aligned}\mathrm{E}\left[\left(u^{(j)}\right)^2\right] &= \mathrm{E}\left[\sum_{m,k} u_m u_k \mathbf{1}_{B_j}(\mathrm{X}_m)\mathbf{1}_{B_j}(\mathrm{X}_k)\right] \\ &= \sum_{m\neq k} u_m u_k P_m^{(j)} P_k^{(j)} + \sum_m u_m^2 P_m^{(j)} \\ &= \left(\mathrm{E}\left[u^{(j)}\right]\right)^2 + \sum_m u_m^2 \left(P_m^{(j)} - \left(P_m^{(j)}\right)^2\right)\,.\end{aligned}$$

Note that $P_m^{(j)} \in \{0, 1\}$ if the mth member lies always in the same bin. The lesser the variability of each member, the lesser the bias (like for the ensemble CRPS bias).

Consequently, each term $\mathrm{E}\left[\left(u^{(j)}\right)^2\right]$ generates an additional $\sum_m u_m^2 \left(P_m^{(j)} - \left(P_m^{(j)}\right)^2\right)$ in Equation 4.22. They accumulate to the bias of the score, which is equal to

$$\sum_m u_m^2 \sum_{j=0}^{S} \frac{P_m^{(j)} - \left(P_m^{(j)}\right)^2}{\mathrm{F}_{j+1} - \mathrm{F}_j} \,.$$

Unbiased estimates of the score can be delivered by counteracting the bias. An interpretation of the bias counteraction with classes of i.i.d. members is left for future research.

5 Application of online CRPS learning to probabilistic PV power forecasting

We provide probabilistic forecasts of photovoltaic (PV) production, for several PV plants located in France up to 6 days of lead time, with a 30-min *timestep. First, we derive multiple forecasts from numerical weather predictions (ECMWF and Météo France), including ensemble forecasts. Second, our parameter-free online learning technique generates a weighted combination of the production forecasts for each PV plant. The weights are computed sequentially before each forecast using only past information. Our strategy is to minimize the Continuous Ranked Probability Score (CRPS). We show that our technique provides forecast improvements for both deterministic and probabilistic evaluation tools.*

This chapter is a research paper written with Christophe Chaussin and Vivien Mallet, and submitted to International Journal of Forecasting.

Contents

Introduction

Improved photovoltaic power integration needs better power forecasts. Forecasters may pursue efforts to improve meteorological models, weather-based power models or statistical post-processing methods. For our part, we focus on the following case: a forecaster, willing to provide probabilistic PV power forecasts, retrieves multiple meteorological forecasts (possibly from various sources). In this general setting, numerous state-of-the-art methods can be tested and combined.

Meteorological forecasts can either be deterministic single forecasts or an ensemble of forecasts, usually at coarser resolution. Inman et al. [IPC13] provide a review of PV forecasting methods with deterministic forecasts. Ensemble forecasting and more generally probabilistic forecasting has been widely covered in the meteorological community [GK14]. Only recently, ensemble-based forecasting techniques are tested for PV [Zam+14], while these techniques are more common for wind and wind power forecasting [RSS15].

A recent benchmark of deterministic and probabilistic PV forecasts is analyzed in Sperati et al. [Spe+15], along with classical diagnostic tools. Probabilistic forecasts rely on the estimation of quantiles of the predicted probability density function (PDF). Quantile regression [APN15] and analogs [Ale+15; HP15] are amongst most popular techniques for quantile estimation in PV. These techniques do not require an ensemble of forecasts as they can rely only on the historical variability of the forecasts and production data.

A forecaster having multiple forecasts hopefully wishes to combine them in an optimal way. Online learning techniques provide rules for combining forecasts, see the monograph Cesa-Bianchi and Lugosi [CL06]. The combination rules stemming from online learning depend only on the available past information at each forecast step and come with theoretical performance guarantee under essentially no assumptions (concerning prior weights, underlying stochastic process or distributions). These techniques have been tested for several applications: electricity consumption, ozone concentration, wind and geopotential fields, and solar irradiance [Sto10; MSM09; Mal10; Dev+13; Bau15; Tho+15].

This paper presents application results with our innovative approach [TMB16], whose purpose is to combine multiple forecasters in a linear opinion pool [GM90; GA11]. The originality of our technique is to use combination rules deriving from online learning techniques in order to minimize the CRPS of the weighted empirical distribution function. We stress here the fact that our method provides theoretical guarantee and that it does not rely on distribution assumptions. Besides, the algorithm has a low computational cost and is parameter-free. Our framework is inspired from the work of Gaillard et al. [GGN16], which focuses on quantile scoring functions.

Minimizing the CRPS is a common strategy in the meteorological literature to obtain calibrated probabilistic forecasts. However, standard techniques do not offer theoretical guarantees of robustness and usually resort to strong assumptions on the distributions. For example, Bayesian model averaging (BMA) techniques provide a mixture of parametric distributions, usually a Gaussian sum [Gne+05] or gamma distributions sum for wind and precipitation applications [SGR10; Slo+07]. Non-homogeneous regression fits

Label	Nature	Origin	Timestep	Resolution	Base time	# of forecasts
HRES	deterministic	ECMWF	3 h	0.13°	0 h	1
ARPEGE	deterministic	Météo France	1 h	0.10°	0 h	1
ENS	ensemble	ECMWF	3 h	0.25°	0 h	50
PEARP	ensemble	Météo France	3 h	0.20°	18 h	34

Table 5.1 – Forecast weather data. The indicated resolutions may change for further lead times than those of the present article.

D	D + 1	D + 2	D + 3	D + 4	D + 5
PEARP	PEARP	x	x	x	x
Det	Det	Det	Det	x	x
ENS	ENS	ENS	ENS	ENS	ENS

Table 5.2 – Forecast availability with lead time. PEARP is the Météo France ensemble, Det defines the deterministic forecasts Arpège and HRES, and ENS is the ECMWF ensemble.

the parameters of a parameterized distribution using characteristics of the ensemble of forecasts [Gne+05; Wil09; TG10]. For instance, a Gaussian distribution is fitted using a linear model between the mean of the distribution and the mean of the forecasts. Besides, likelihood maximization with the logarithm loss is not an appropriate tool in our setting since it fails to produce satisfactory scores for a mixture of Dirac distributions. A discussion on local scores such as the logarithm loss is adressed by Bröcker and Smith [BS07b].

In Section 5.1, we introduce the production data sets and the forecasts from ECMWF and Météo France. We also detail our method to generate PV forecasts from meteorological data. The evaluation tools are described in Section 5.2. Our statistical post-processing method is explained in Section 5.3. Numerical results and discussions are developed in Section 5.4.

5.1 Methods

5.1.1 Production and meteorological data

The production data cover 219 PV power plants in metropolitan France with 21 consecutive months (January 2012 to October 2013). The total power of the plants is referred to as France production. We wish to provide production forecasts for each power plant and for France production. The data are shown as load factor, i.e. scaled by the installed capacity. France production forecasts are the weighted sums of the plant forecasts w.r.t. the installed capacity of each plant.

Forecast data are summarized in Table 5.1 and 5.2. We use data from two meteorological centers (ECMWF and Météo France), both deterministic forecasts and ensembles of forecasts: HRES and ENS for ECMWF, and ARPEGE and PEARP for Météo France [Cou+91; Des+15; Pal+09], up to a lead time of 6 days. Note that the deterministic forecasts are not the unperturbed members of the ensembles of forecasts

but different forecasts, with better resolution.

We are interested in predicting the 30-min average power output of the plants. We only show results for the following hours of the day 0600, 0900, 1200, 1500, and 1800 (where 0600 refers to 6:00 UTC) in order to save computation time and to avoid the issue of temporal interpolation of our forecast data. Ensemble forecasts are only solar irradiance forecasts while deterministic forecasts also include total cloud cover and 2-m temperature. The ensemble PEARP is available for longer lead times but only with a time step of 6 (and not 3) hours. Consequently, for our application we restricted the use of PEARP up to 2 days.

5.1.2 Conversion of meteorological forecasts to production forecasts

Our regression technique is inspired from Bacher et al. [BMN09] and Lorenz et al. [Lor+09b]. This regression technique has been successfully applied in the benchmark of Sperati et al. [Spe+15], where the technique ranked first in the deterministic PV forecasting competition. For a given deterministic forecast, the following technique is applied for each time of the day and for each power plant independently. The training set ranges from early 2012 to February 2013 (nearly 400 days). The testing set with the remaining days of 2013 is about 240-day long.

First, clear sky indices τ_P and τ_I are generated from the production P and the solar forecasts I:

$$\tau_P = \frac{P}{P_{cc}} \quad \text{and} \quad \tau_I = \frac{I}{I_{cc}}\,, \tag{5.1}$$

where clear sky production P_{cc} and clear sky solar radiation I_{cc} are the production and solar radiation in clear sky conditions. The clear sky profiles P_{cc} and I_{cc} are respectively estimated from the production P and the solar forecasts I thanks to quantile regression introduced in Section 5.1.3.

The core of the statistical analysis is a linear regression between the production index τ_P and the meteorological variables (the clear sky index τ_I, the total cloud cover Tcc, and the temperature T_{2m}). Non-linear dependencies are taken into account by introducing several terms such as the squared clear sky index τ_I^2 and cross terms between variables $\tau_I(T_{2m} - \overline{T}_{2m})$. The quantity $T_{2m} - \overline{T}_{2m}$ is the deviation of the temperature T_{2m} from its local average value $\overline{T}_{2m}$. The linear regression estimates the coefficients a_i to produce

$$\widehat{\tau_P} = a_0 + a_1\tau_I + a_2\tau_I^2 + a_3 Tcc + a_4\tau_I(T_{2m} - \overline{T}_{2m})\,. \tag{5.2}$$

A secondary statistical model is then fitted on the residuals $\widehat{\tau_P} - \tau_P$. The objective of this secondary model is to reduce the seasonal biases and other remaining errors of the first model. We use the elapsed time s from January 1^{st} of the current year, and the production forecasts $\widehat{\tau_P}$ as inputs to build

$$\widehat{\widehat{\tau_P}} = b_0\widehat{\tau_P} + \sum_\lambda b_\lambda \sin(\lambda s) + b'_\lambda \cos(\lambda s)\,, \tag{5.3}$$

a linear additive model on $\widehat{\tau_P}$ and trigonometric polynomials of s, see Lorenz et al. [Lor+09b].

The model parameters of the first two steps are set with the forecasts whose lead times are inferior to 24 h. The motivation behind is that the forecasts with short lead times are presumably the most accurate forecasts to fit the main parameters. Still, forecasting long lead times may necessitate slight corrections compared to short lead times, hence we introduce a third step, which takes into account the lead time of the forecasts. The third step is a multiplicative correction λ applied to $\widehat{\widehat{\tau_P}}$, such that $\lambda\widehat{\widehat{\tau_P}}/\tau_P$ is equal to 1 on average.

The statistical regression scheme is slightly different for ensemble and deterministic forecasts. For ensemble forecasts, the input variable of the linear regression is simply the solar irradiance of the unperturbed member without other weather variables. The same conversion model is used for all the members of a given ensemble.

5.1.3 Quantile forecasts

For each deterministic production forecast, we build 19 quantile forecasts (of order 5 to 95) for a total of 38 additional forecasts. They are referred to as deterministic quantiles as opposed to the ensemble members. The idea is to train (on the training data set) a pick-up rule that gives production quantile forecasts according to the value of the deterministic production forecast, see [NMN06]. We follow the idea that we should first precisely estimate the mean of the distribution and only then estimate the quantiles.

Quantile regression uses a piecewise linear asymmetric loss function $\mathrm{QS}_\alpha(x, y)$ called the quantile score (or pinball loss) of level α [KH01]:

$$\mathrm{QS}_\alpha(x, y) = \alpha(y - x)_+ + (1 - \alpha)(x - y)_+ \,, \tag{5.4}$$

where $(\cdot)_+ = \max(\cdot, 0)$. The expectation (over y) of $\mathrm{QS}_\alpha(x, y)$ is minimized if x correctly estimates the quantile of level α of y.

We apply quantile regressions on the residuals of the deterministic forecast obtained at the end of Section 5.1.2. The inputs are the deterministic forecast and trigonometric polynomials of s, similarly to the seasonal bias reduction of the deterministic forecasts (step 2 in Section 5.1.2). The quantile regressions are carried out independently for each lead time.

The clear sky profiles P_{cc} and I_{cc} are also based on quantile regressions but with only trigonometric polynomials of s as inputs. The chosen levels of quantiles to build clear sky profiles are close to 90%.

Concerning France production deterministic quantiles, they are not set to a weighted sum of quantiles of the plants, but they are determined from the deterministic forecast of France production. In other words, the deterministic forecasts of the plants are summed to generate France production forecast, and this forecast is used to generate quantile forecasts for France production.

At this point, the forecaster has a total of $50 + 34 + 2 \times 19 + 2 = 124$ forecasts up to the lead time of 48 hours, 90 forecasts up to the lead time of 96 hours, and 50 forecasts up to the lead time of 138 hours.

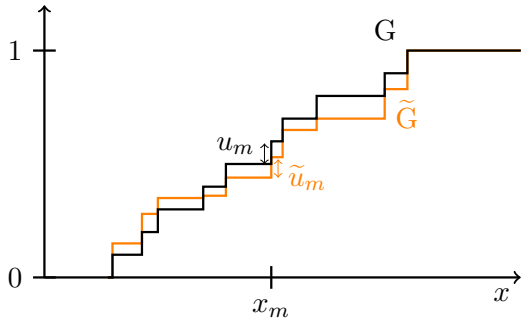


Figure 5.1 – Illustration of weighted CDFs. The CDFs G and $\widetilde{\mathrm{G}}$ are built with the same locations x_m. However the weight u_m or $\widetilde{u}_m$ given to a member is different for G and $\widetilde{\mathrm{G}}$.

5.1.4 Linear opinion pools

Let the x_m be M forecasts (or members). The unit Cumulative Distribution Function (CDF) $\mathrm{H}_m(x) = \mathrm{H}(x - x_m)$ equals 0 before x_m and 1 otherwise. The forecaster's CDF $\mathrm{G} = \sum_m u_m \mathrm{H}_m$ is designed as a weighted combination of unit step functions. The mth step of G is centered on x_m and its height equals the weight u_m. The weights u_m are non-negative and sum to one ($\boldsymbol{u}$ in $\mathcal{P}_M$ the simplex of $\mathbb{R}^M$). This weighted CDF is also known as model mixture or linear opinion pool. Using a discrete CDF based on several forecasts allows us to model any CDF without distribution assumption.

The impact of the weights u_m are illustrated in Figure 5.1 and 5.2. Two CDFs G and $\widetilde{\mathrm{G}}$ using the same locations x_m are shown in Figure 5.1. The CDF G is built with uniform weights $u_m = 1/M$, while the weights $\widetilde{u}_m$ of $\widetilde{\mathrm{G}}$ are not uniform. We show in Figure 5.2 an illustration of probabilistic forecasts in two different cases: with equal weights for all members and with possibly different weights given by our online learning algorithm. A visual inspection indicates that the online learning algorithm provides a better estimation of the median and a larger spread of the distribution. We emphasize that methods involving weighted empirical distribution functions necessitate that the forecasts x_m are sufficiently dispersed.

5.2 Evaluation

In the following we describe classical diagnostic tools used in Section 5.4, see for example the monograph of Jolliffe and Stephenson [JS12] for further references.

We begin by describing the CRPS as it is at the heart of our learning method.

5.2.1 The CRPS

The CRPS is a classical scoring function in meteorology [Her00; CT05]. The CRPS is the generalization over all thresholds of the Brier score [Bri50]. Let G be the cumulative distribution function of a forecaster describing the i.i.d. random variables X and X′, and y be the observation revealed to the forecaster. The CRPS is defined as

$$\mathrm{CRPS}(\mathrm{G}, y) = \int (\mathrm{G} - \mathrm{H}_y)^2 = \mathrm{E}(|\mathrm{X} - y|) - \frac{1}{2}\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)\,, \tag{5.5}$$

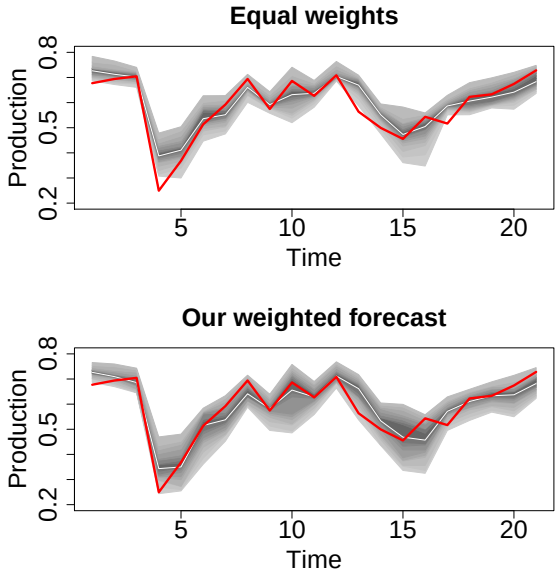


Figure 5.2 – Time series of France production forecasts scaled by the installed capacity (12 hours of lead time, for several consecutive days). Top: equal weights for all members, (b): our forecast with online learning of the weights. Real production is in red and the median of the forecasted distribution is in white.

where $\mathrm{H}_y(x) = \mathrm{H}(x - y)$ is the CDF assigned to y, the unit step function H centered on y. The CRPS reduces to the absolute error for deterministic forecasts.

Assuming that y is a random variable, described by the CDF F, the averaged quantity $\mathrm{E}_y(\mathrm{CRPS}(\mathrm{G}, y))$ (on the observation) is minimized only for $\mathrm{F} = \mathrm{G}$. This property makes the CRPS a strictly proper scoring rule [GR07], and as such it explains why the CRPS is a classical evaluation tool for probabilistic forecasts.

We highlight the fact the CRPS can also be written as a sum of quantile scores [GR11]:

$$\mathrm{CRPS}(\mathrm{G}, y) = 2 \int_0^1 \mathrm{QS}_\alpha(\mathrm{G}^{-1}(\alpha), y) \mathrm{d}\alpha \,. \tag{5.6}$$

The strategies of minimizing the CRPS or minimizing several quantile losses are therefore closely related.

For a CDF step function, the corresponding CRPS is computed as:

$$\mathrm{CRPS}\left(\sum_{m=1}^{M} u_m \mathrm{H}_m, y\right) = \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2} \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| \,. \tag{5.7}$$

which is also concisely noted $\ell(\boldsymbol{u})$ in Section 5.3.

5.2.2 Other diagnostic tools

The scores are all presented only for the test period, usually averaged over time. Besides the CRPS, we also show results for the celebrated RMSE and MAE for which

our forecast is the weighted average $\sum_{m=1}^{M} u_m x_m$. The RMSE of the predictions $\widehat{y}$ with respect to the observations y is given by

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\widehat{y}_t - y_t)^2}\,, \tag{5.8}$$

and for the MAE:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\widehat{y}_t - y_t|\,. \tag{5.9}$$

We use daily scores to show the deterioration of the scores with the increasing lead time. To keep the range of the daily score consistent, the daily score is weighted by the average production of the related hour of the day $\bar{y}_h$. For a score S_h depending on the lead time h, the daily score

$$S^{(d)} = \frac{\sum\limits_{h} S_h \times \bar{y}_h}{\sum\limits_{h} \bar{y}_h} \tag{5.10}$$

is computed with summation over the available lead times h corresponding to the same daily lead time.

Skill scores are useful to compare prediction performance. In this paper, the reference prediction chosen for skill scores is our forecast. Skill scores for a given score S are written

$$S_{pred}^{skill} = \frac{S_{ref} - S_{pred}}{S_{ref}}\,, \tag{5.11}$$

so that our forecast shows better scores when the skill scores of the other forecasts are negative.

5.3 Online learning with the CRPS

5.3.1 Background

Our objective is to produce an optimal combination (of step functions), or more precisely, to minimize the regret due to unavoidable loss w.r.t. the best learning algorithm for a given class of algorithms. A learning algorithm, along with its weight update rule, is used to find weights for each time step using only available past information. In other words, we want to use an update rule that indicates the value of the weights $u_{m,t}$ and relies only on the values of the past forecasts and observations $x_{m,t'}$ and $y_{t'}$ with $t' < t$.

Online learning algorithms come with a theoretical guarantee of long term performance. The guarantee is often expressed under the form of a regret bound:

$$\sup\left[\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}_M}\sum_{t=1}^{T}\ell_t(\boldsymbol{u})\right] \leqslant o(T)\,, \tag{5.12}$$

where the supremum is taken over all possible values of $x_{m,t}$ and y_t. The notation ℓ_t refers to the CRPS as in Equation 5.7 with a highlighted dependency on the weight.

In the sense of the theoretical guarantee, our algorithm competes against the best combination with weights constant in time, which can be known only at the end of the experiment and is called the oracle. By definition, the oracle has a better score than individual forecasts (MAE of each forecast), and than any subset ensemble with uniform weights.

It is common practice in online learning to use linearized losses, by computing the loss gradients w.r.t. the weights. For the CRPS, the loss gradient $\widetilde{\ell}_{m,t}$ of the mth forecaster can be written as

$$\widetilde{\ell}_{m,t} = \frac{\partial \ell_t}{\partial u_m}(\boldsymbol{u}_t) = |x_{m,t} - y_t| - \sum_{k=1}^{M} u_{k,t}|x_{m,t} - x_{k,t}| + y_t - \sum_{k=1}^{M} u_{k,t}x_{k,t}\,. \tag{5.13}$$

The last two terms are identical for all forecasters and appear due to terms $1-\sum_{m=1}^{M} u_m$ hidden in the expression of the CRPS, see Appendix B of Thorey et al. [TMB16]. The loss gradient is balanced between the distance of $x_{m,t}$ to y_t and the weighted distance of $x_{m,t}$ to the ensemble members. A very good member is therefore close to the observation and far from the other members. A neutral member is equally distant to the observation and the other members.

The loss linearization shifts a regret against the best combination as Equation 5.12 to a regret against the best member only, as we now detail (see also Devaine et al. [Dev+13]). The convexity and the differentiability of ℓ_t gives

$$\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \leqslant (\boldsymbol{u}_t - \boldsymbol{u})^\top \nabla \boldsymbol{\ell}_t(\boldsymbol{u}_t) = \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \boldsymbol{u}^\top \widetilde{\boldsymbol{\ell}}_t\,. \tag{5.14}$$

for any two vectors $\boldsymbol{u}_t, \boldsymbol{u} \in \mathcal{P}_M$. Summing over time, we get the following regret bound inequalities:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) = \sup_{\boldsymbol{u}\in\mathcal{P}} \left(\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \right) \tag{5.15}$$

$$\leqslant \sup_{\boldsymbol{u}\in\mathcal{P}} \left(\sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \boldsymbol{u}^\top \widetilde{\boldsymbol{\ell}}_t \right) \tag{5.16}$$

$$= \sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \min_{\text{expert k}} \sum_{t=1}^{T} \widetilde{\ell}_{k,t}\,. \tag{5.17}$$

As a consequence, an algorithm formulated for linear losses $\sum_{m=1}^{M} u_m \widetilde{\ell}_m$ and coming with theoretical guarantee (on the expression 5.17) may be used with any convex differentiable loss, by applying the algorithm on the gradient losses. A theoretical guarantee for the non linear losses $\ell_t(\boldsymbol{u})$ is then obtained. In other word, knowing a regret bound for expression 5.17 provides a regret bound for expression 5.15.

5.3.2 Example of general algorithm

Initialization: $\boldsymbol{u}_1$;

For each time index t = 1, 2, ..., T

1. get the vector of predictions data $\boldsymbol{x}_t$,

update the learning rate of each member	$\eta_{m,t} = 1/\left(1 + \sum_{t'=1}^{t} (\boldsymbol{u}_{t'}^\top \widetilde{\boldsymbol{\ell}}_{t'} - \widetilde{\ell}_{m,t'})^2\right)$
update the regret of each member	$R_{m,t} = R_{m,t-1} + \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$
compute the weights	$u_{m,t+1} = \eta_{m,t}(R_{m,t})_+ \;/\; \boldsymbol{\eta_t}^\top (\boldsymbol{R}_t)_+$

Table 5.3 – ML-Poly algorithm, at time t after y_t is given. The vectors $\boldsymbol{\eta_t}$ and $\boldsymbol{R}_t$ have M coordinates, respectively $\eta_{m,t}$ and $R_{m,t}$. The functions $(\cdot)_+$ applied to a vector are applied to all the vector's components.

2. compute the forecaster's choice G_t with $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$,
3. get the verification y_t and compute $\boldsymbol{u}_{t+1}$, based on the update rule.

The initial weight vector $\boldsymbol{u}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$.

5.3.3 ML-Poly

In this article we use a learning algorithm from [GSE14] called ML-Poly for Polynomially weighted averages with multiple learning rates. The algorithm ML-Poly, described in Table 5.3, has no parameters. The algorithm relies on terms $\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$ that compare the performance of each member to the performance of the weighted ensemble. The learning rate $\eta_{m,t}$ checks whether a forecaster's performance is in average close to the performance of the weighted forecast, and the regret $R_{m,t}$ quantifies the regret for not having given higher weights to a forecaster. The ideas behind ML-Poly are on the one hand an adaptation of the algorithm Prod of Cesa-Bianchi et al. [CMS05] to multiple learning rates, and on the other hand the introduction of the polynomial potential described in Cesa-Bianchi and Lugosi [CL03] and giving the terms $(R_{m,t})_+$.

The regret bound of ML-Poly is expressed against the best member for the linearized losses. For all sequences of losses $\widetilde{\ell}_{m,t} \in [0,1]$, the cumulated loss of ML-Poly is bounded:

$$\sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t \leqslant \min_{1 \leqslant m \leqslant M} \left\{ \sum_{t=1}^{T} \widetilde{\ell}_{m,t} + \sqrt{M(1+\ln(1+T)) \left(1 + \sum_{t=1}^{T} (\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2\right)} \right\}. \tag{5.18}$$

As opposed to the bound of Equation 5.12, the bound of ML-Poly is of second order due to the term $\sum_{t=1}^{T} (\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2$. The worst case scenario gives a bound $\mathcal{O}(\sqrt{MT \ln T})$, indicating that even in the worst case, the weighted forecast will perform at least as well as the best forecast. In the case of i.i.d. sequences of losses, the regret bound is practically constant. A detailed analysis of second-order bounds can be found in [GSE14]. Besides, other algorithms showing second order bounds are described in Koolen and Van Erven [KV15], Luo and Schapire [LS15], and Wintenberger [Win17].

5.4 Application

5.4.1 Experiment setup

Local production data may be unfortunately unavailable for given days and plants. In such cases we removed the related data. However we did not modify France production

capacity factor to account for local unavailability, because in our opinion, a challenging task for online learning technique is to reduce biases which may be caused by local null production.

The algorithm is run independently for each lead time and production site (including France production). We run the algorithm as if production data is available at the end of each day. For long lead times h where several observations arrive between the delivery of a forecast and the reception of the corresponding observation, we use shifted weights. At time t we compute u_{t+h} to predict y_{t+h} by using the weights u_{t+h-1} instead of u_t in Table 5.3. For example with the shorthand notation $\boldsymbol{u}_{\text{lead time, day}}$ and with a lead time of 36 h, the weights $\boldsymbol{u}_{36\,\text{h},d}$ were delivered at $d-1$ to forecast $y_{12:00,d}$. The weights $\boldsymbol{u}_{36\,\text{h},d}$ are updated to $\boldsymbol{u}_{36\,\text{h},d+1}$ after $y_{12:00,d-1}$ is known at the end of $d-1$. The key point is that the weight update uses $\boldsymbol{u}_{36\,\text{h},d}$ instead of $\boldsymbol{u}_{36\,\text{h},d-1}$ to check the combination performance against $y_{12:00,d-1}$.

The production forecasts from PEARP and ENS are sorted by rank in order to associate clearly a weight with an ensemble member. As a result, all the members belong to one of the four sorted subensembles, except for the two deterministic forecasts.

We define a climatological reference for diagnostic purposes, called climatology forecast. For time t, we use 2 months of production data centered on t to estimate a so-called climatological mean and 19 quantiles of climatological production. The climatological mean is used for deterministic evaluations (bias, RMSE, MAE) and the quantiles are used for the CRPS. This method produces a rather 'skilled' reference because the climatology is not only evaluated on the training period but on a rolling period.

We define the raw forecast as the forecast with uniform weights. We use this forecast to assess the gain brought by our online learning algorithm.

The results are shown for PV production forecasts only, and not meteorological variables.

5.4.2 Results

In this section we only show the results for the individual plants. The results obtained for France production are quite similar to those obtained for the plants and are shown in Appendix 5.A.

Scores and skill scores

First we show the classical scores RMSE, MAE, CRPS and bias in Figure 5.3 on a daily average, see Equation 5.10. The confidence intervals indicate the variability of the scores obtained for the plants. The scores are shown for our weighted forecast as well as the raw forecast, the ECMWF deterministic forecast and the climatology forecast. Our weighted forecast gets the best scores up to a lead time of 4 days. Note that our forecast has a quite low bias. For days 5 and 6 the ENS members are the only members available in our study, hence a change of slope in the daily scores. Even for a lead time of 6 days, the climatological forecasts is the worst forecast. Therefore numerical weather predictions may be used to forecast plants productions for a lead time of several days at the 30-min timestep. It is noticeable that our regression scheme

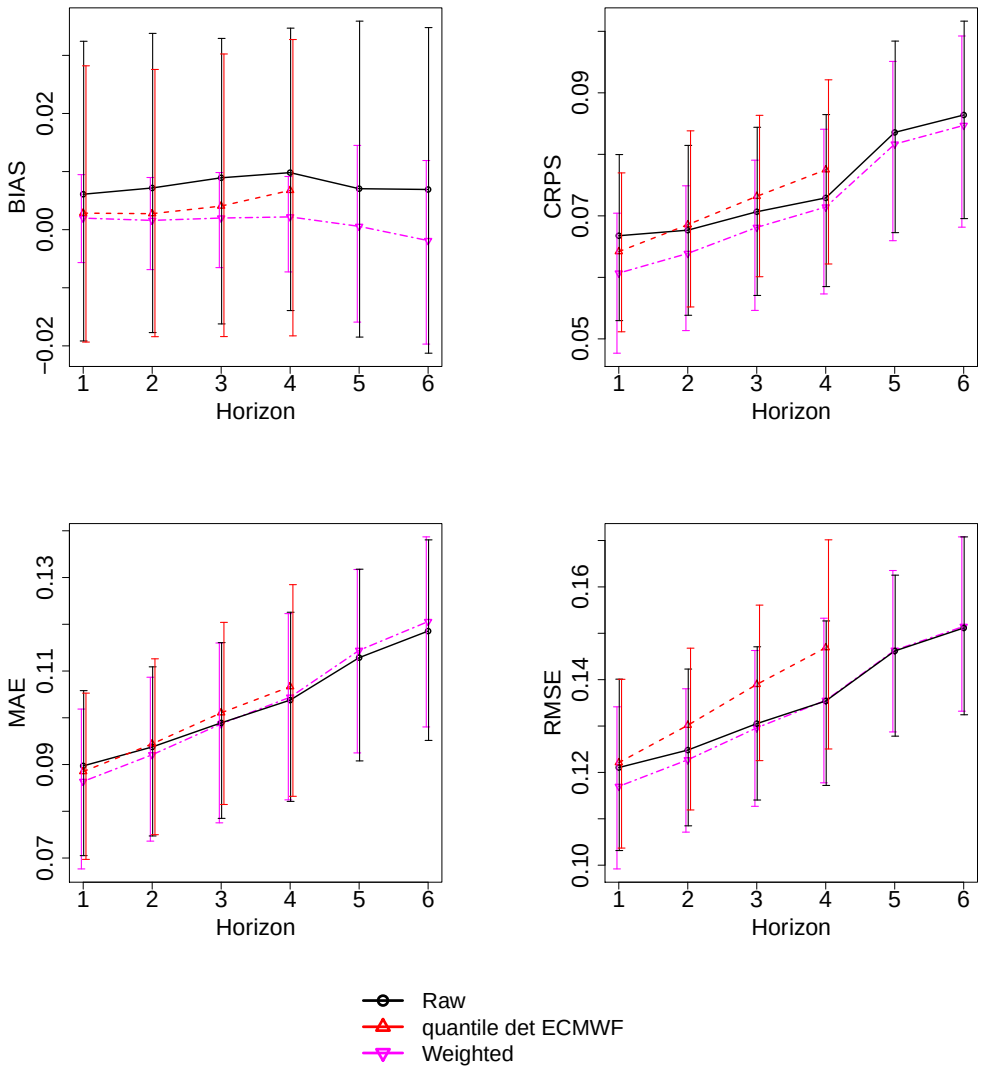


Figure 5.3 – RMSE, MAE, CRPS and bias for the daily scores, for all sites. The results are shown for 3 forecasts: our weighted forecast, the raw forecast (all members with uniform weights), the deterministic forecast of the ECMWF (and its quantiles for the CRPS). The climatology scores are the following : bias = −0.001, CRPS = 0.089, MAE = 0.139, RMSE = 0.167. The confidence intervals are derived from the scores of all sites.

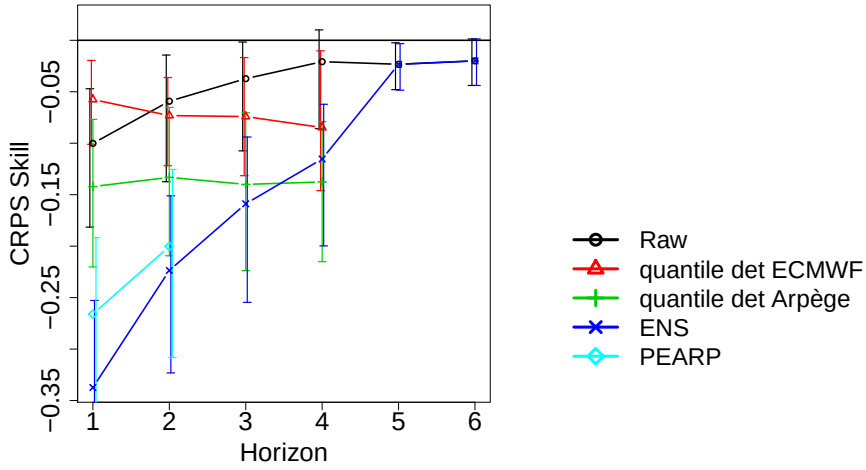


Figure 5.4 – CRPS skill scores for all sites based on Equation 5.11. The confidence intervals correspond to the scores of all sites.

minimizes the CRPS and also achieves improvements on the other scores (RMSE, MAE and bias). We tried to identify situations where our algorithm provides a particular improvement over the raw ensemble, but we did not find discriminatory criteria. For example, the installed capacity of the plant or the CRPS of the raw ensemble are not explanatory statistics of the CRPS skill scores of the raw ensemble.

The CRPS skill scores of 5 ensembles (with uniform weights) are shown in Figure 5.4. The skill scores are assessed against our weighted forecast. The 5 ensembles are the 4 subensembles of our complete ensemble and the complete ensemble as well. Our weighted forecast performs better than any of the 5 ensembles. The best ensemble with uniform weights is (in average) the complete ensemble. This may be due to the variety of the forecasts in the complete ensemble. Although the quantile ensemble from HRES (quantile det ECMWF) performs well before 24 hours of lead time, it is beaten by the complete ensemble afterwards and its skill decreases with time. The skill of the ECMWF ensemble (ENS) increases notably with time, from the worst skill for day 1 to a satisfactory skill for day 4.

Diagnostic tools

Improvements are also shown for several other diagnostic tools but only for a lead time of 36 h (1200, D+1) for the sake of brevity. Better results are obtained for shorter lead times and conversely worst results are obtained for longer lead times. By better we mean improvement of our weighted forecast over the raw ensemble.

The spread-skill diagram checks whether the spread of an ensemble (binned into categories) is consistent with the error of the ensemble mean. The squared spread $\sum u_m(x_m - \boldsymbol{u}^\top \boldsymbol{x})^2$ and the square error $(\boldsymbol{u}^\top \boldsymbol{x} - y)^2$ are averaged in each bin and their rooted square are plotted against each other. The spread and the error should be ideally equal [For+14]. On the graph, the curves should match the first diagonal. The spread-skill diagram of the ensembles of our study is shown on Figure 5.5. We see that

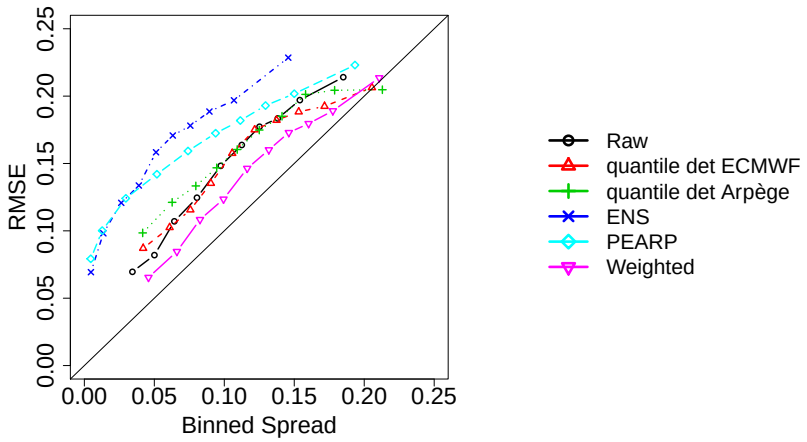


Figure 5.5 – Spread skill diagrams for 36 h of lead times for all sites.

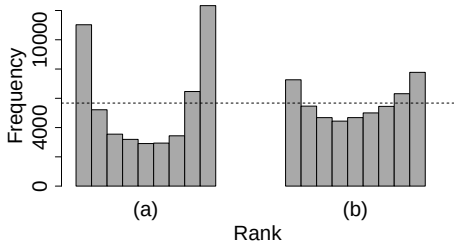


Figure 5.6 – Rank histograms for 36 h of lead times for all sites; (a) the raw ensemble; (b) our weighted forecast. The dotted line illustrates the ideal case of a flat rank histogram.

our weighted forecasts are closer to the first diagonal than any other subensemble with uniform weights. Our weighted forecasts for plants are still under-dispersive, while the correction is better in the case of France production as shown in Appendix 5.A. The weights provided by the online learning algorithm are larger for the outer members of the ensemble, and especially the lowest members. Consequently the spread of the weighted ensemble is larger than the spread of the raw ensemble and the positive bias of the raw ensemble is mitigated. Besides, the ECMWF ensemble shows the lowest spread and the ensemble PEARP presents very large and very small spreads. However, when the ensemble PEARP shows a small spread, the error is quite larger than the spread.

The rank histogram [And96; TVS99; HC97] or Probability Integral Transform (PIT) is built with the values of the CDFs of the forecaster reached by the verifications along an experiment. The ideal rank histogram is flat. The rank histogram of our weighted forecast and the raw ensemble are shown in Figure 5.6. The rank histogram of our weighted forecast is closer to the ideal rank histogram than the rank diagram of the raw ensemble. The raw ensemble is under-dispersive, since it presents a U-shape. This

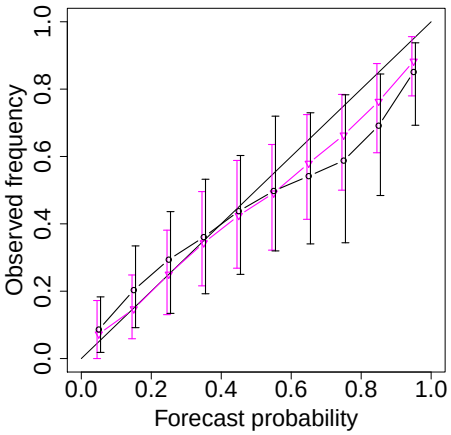


Figure 5.7 – Reliability diagrams for lead times 36 hours for all sites; black circle: the raw ensemble; magenta triangles: our weighted forecast.

is consistent with the results shown on Figure 5.5.

For a given binary event, the reliability diagram checks whether the observed frequency and the forecasted frequency of the event match [Atg04; BS07a]. The forecasted probabilities of the event are binned into categories. The observed frequency of the event for each category is the share of occurence of the event. The ideal reliability diagram shows a curve along the first diagonal. We use the following event "the production level is lower than the average production", where we use the climatological production defined above as local average production. We show the reliability diagrams of our weighted forecast and the raw ensemble in Figure 5.7. We see that our weighted forecast is very well calibrated for event with low probability, but tends to overpredict the occurence of the event when the event is highly likely.

Conclusion

We have applied the algorithm ML-Poly for the minimization of the CRPS, in order to provide probabilistic forecasts. The algorithm does not depend on any parameter or assumptions on distributions such as Gaussianity, and comes with theoretical guarantee of performance. The regret bound ensures our forecast to perform at least as well as the best forecast in the ensemble.

Our case study investigated the PV production of several power plants in France and the total production of the plants. We have shown that our weighted forecast improves on the raw ensemble, which is the best ensemble with uniform weights. Interestingly, we show that CRPS minimization brings improvement on classical scores for the ensemble mean and probabilistic diagnostic tools. Indeed, the forecasting capability measured by classical scores (RMSE, MAE, CRPS and bias) are improved by our online learning algorithm up to a lead time of 4 days. Besides, the online learning algorithm provides a spread correction as shown on the spread-skill diagrams and on the rank histograms.

The results obtained for France production forecasts and plants forecasts are quite similar.

Future work should investigate the generation of specialized experts on meteorological regimes. For example, an expert specialized in clear sky production could improve the forecasting capability of the ensemble. The quantiles are already specialized, but the ensemble members from ENS and PEARP are converted to production using the same model as for the control member. The investigation of weights prior may also be of interest. The update rule ML-Poly does not use the value of the upcoming forecasts $x_{m,t}$ for computation of the weights $u_{m,t}$, while weights prior may take this additional information into account.

Appendix 5.A Results for France production

In this Appendix, we show the results for France production, while the results for the individual sites are shown in Section 5.4. The results of France production forecasts and plants forecasts are roughly similar. Our online learning algorithm provides improvements over the raw ensemble up to a lead time of a few days. Because it is easier to forecast the power output of the total production, the forecast quality is better than for individual sites. This statement is verified for all diagnostic tools shown below.

We show in Figure 5.8 the average bias, CRPS, MAE and RMSE for France production. We see the scores of the sites are more than twice as large as the score of France production, but for the bias. Our online learning algorithm provides improvement for bias, CRPS, MAE and RMSE up to a lead time of 4 days.

The CRPS skill scores are shown in Figure 5.9. Once again, the score trends are mostly equivalent to those obtained for the sites. The quantile ensemble from HRES (quantile det ECMWF) has good scores for short lead times and the raw ensemble is the best ensemble with uniform weights after 24 h of lead time. Our online learning algorithm provides an improvement of roughly 10% over the raw ensemble for the first 24 h of lead time. This improvement decreases with time quickly than for the sites. It is remarkable that the CRPS skill score of the quantile ensemble from Arpège ("quantile det Arpège") shows much better results for the plants than for France production. Indeed the skill score of "quantile det Arpège" is around -15% for the plants and is stable, while it is at least below -24% for France production. For days 5 and 6, the weights brought by our algorithm do not vary much from the uniform distribution. Consequently, the skill scores are close to one.

The following probabilistic diagnostic tools are only for 0900, 1200, and 1500 of day 2 (lead times 33, 36, 39 hours). In Figure 5.10, we compare the rank histogram for the raw ensemble and our weighted forecast. The raw ensemble is largely under-dispersive with a positive bias (over-estimation). Our online learning algorithm manages to reduce the under-dispersion of the raw ensemble. This statement is verified on the spread skill diagram in Figure 5.11. We see that the spread and errors of our weighted forecasts match approximately, while the other ensembles (with uniform weights) are under-dispersive with respect to their errors.

A correction of the forecast reliability is illustrated in Figure 5.12. The event "the production level is lower than the average production" is used (same as for the sites).

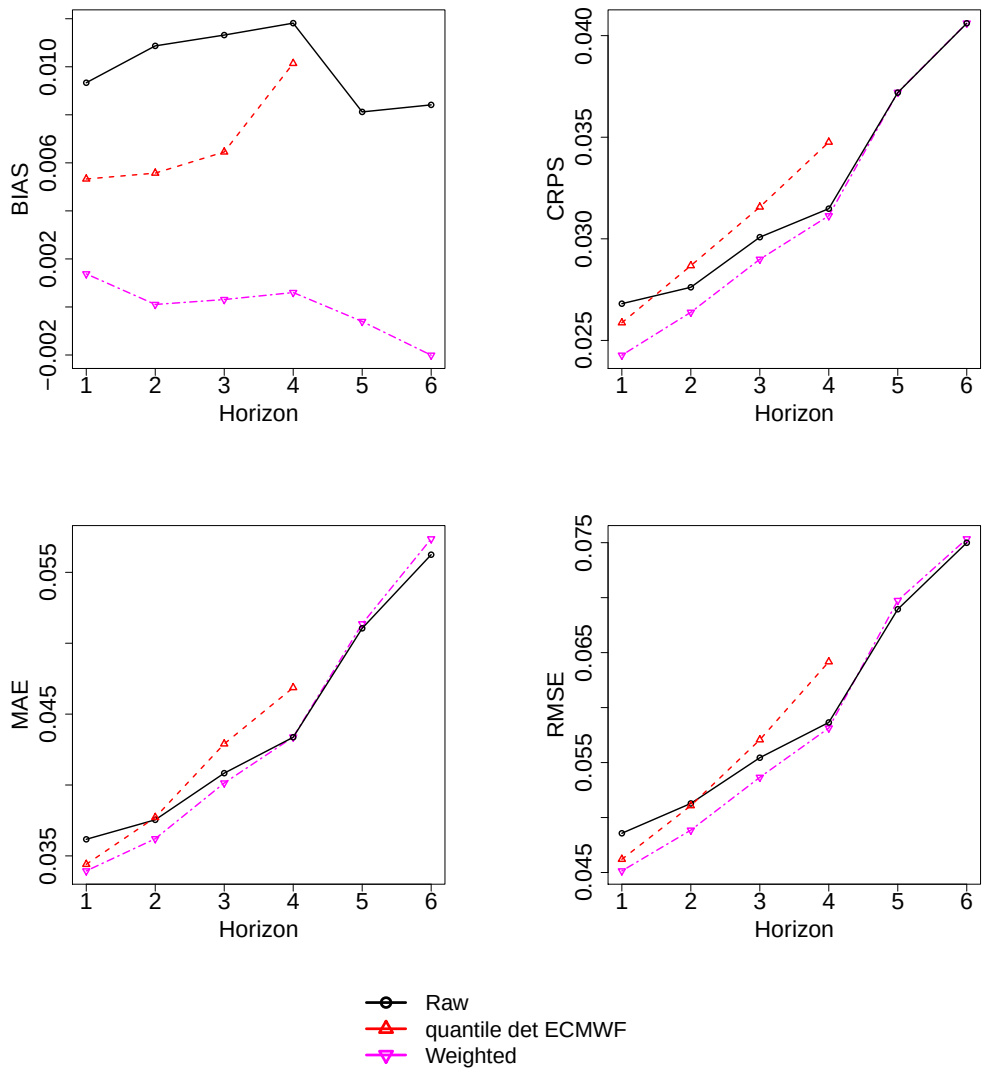


Figure 5.8 – RMSE, MAE, CRPS and bias for the daily scores of France production. The results are shown for 3 forecasts: our weighted forecast, the raw forecast (all members with uniform weights), the deterministic forecast of the ECMWF (and its quantiles for the CRPS). The climatology scores are the following : bias = −0.001, CRPS = 0.055, MAE = 0.081, RMSE = 0.101.

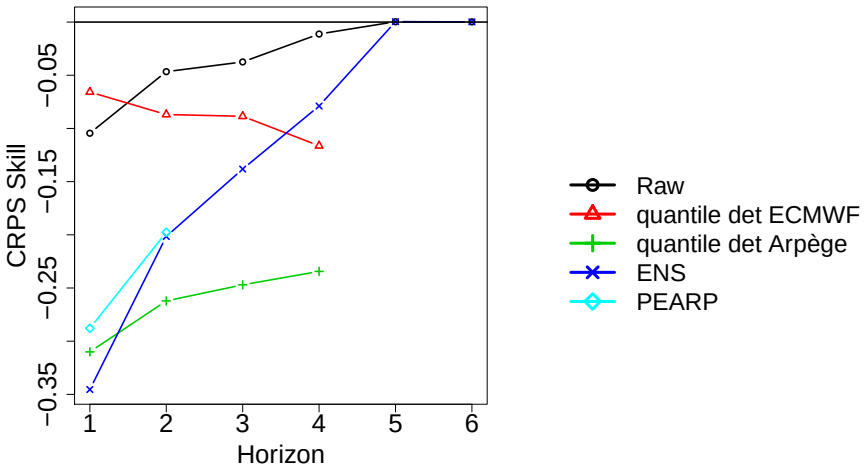


Figure 5.9 – CRPS Skill score.

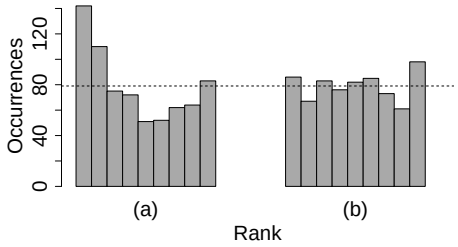


Figure 5.10 – Rank histograms for lead times 33, 36, 39 hours; (a) the raw ensemble; (b) our weighted forecast. The dotted line illustrates the ideal case of a flat rank histogram.

A visual inspection shows that the raw ensemble tends to underpredict the occurence of low production for a forecasted frequency between 0.3 and 0.7, when the event is likely to occur. Our weighted forecast does not show this tendency and is symmetrical with respect to the first diagonal although not perfectly aligned.

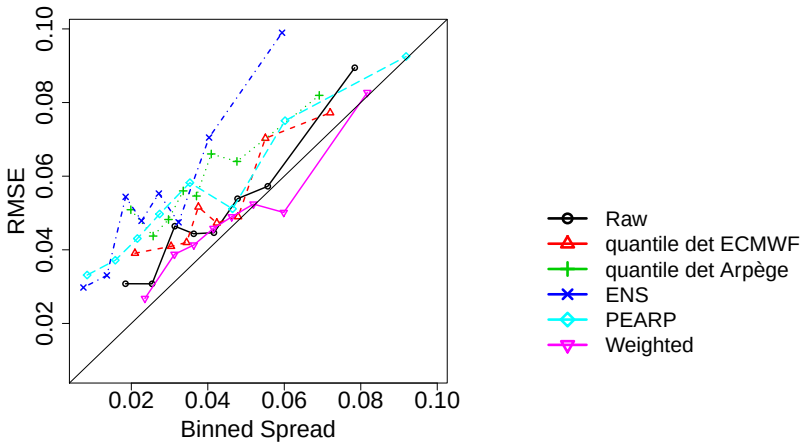


Figure 5.11 – Spread skill for lead times 33, 36, 39 hours.

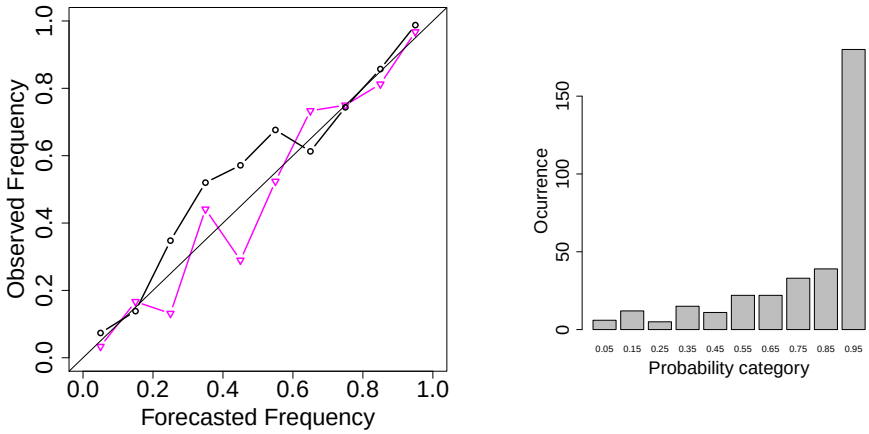


Figure 5.12 – Reliability diagrams for lead times 33, 36, 39 hours; black round: the raw ensemble; magenta triangles: our weighted forecast.

6 PV probabilistic forecasts with the AROME high resolution forecasts

The high-resolution forecasting system AROME delivers dense spatio-temporal information for short lead times. In this chapter, we generate numerous PV forecasts using AROME solar radiation forecasts by leveraging multiple statistical models and exploiting the available spatio-temporal information. These forecasts are then combined with online learning techniques. We study the calibration of the resulting forecasts and improve the calibration thanks to additional quantile predictions. Finally, AROME forecasts are combined with other forecasts from Météo France and ECMWF.

The work was jointly carried out with Clément Dolou, whose internship at EDF R&D was supervised by the author of this PhD thesis and Christophe Chaussin.

Contents

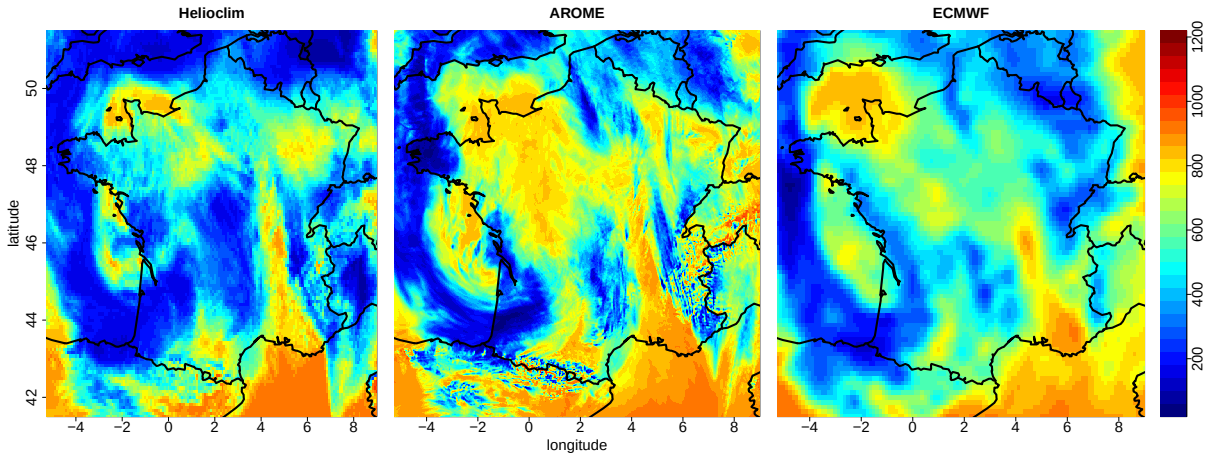


Figure 6.1 – Maps of solar radiation (in $\mathrm{W\,m^{-2}}$) for 2013-06-13 at 12:00 (UTC): HelioClim real-time estimation (left), AROME forecast (center), ECMWF HRES forecast (right).

Besides Arpège, Météo France also provides for France the local model AROME, designed to forecast severe weather events such as heavy rains in the south of France [Sei+11]. AROME forecasts are operational since the end of 2008. Examples of solar radiation maps are provided in Figure 6.1 for satellite estimations HelioClim, AROME and the ECMWF deterministic forecast called HRES. We clearly see the differences in resolution in the three maps. We recall that HRES data have a 3-h resolution while AROME and HelioClim maps are shown here with a 1-h resolution. The weather forecasting system AROME provides high-resolution forecasts with a 2.5-km spatial resolution for our period of study in 2012-2013. In this chapter, we study how AROME forecasts can be used to provide probabilistic forecasts of PV power. This question is of interest for us for the following reasons. The high-resolution of AROME provides rich spatio-temporal information. How can we exploit it? AROME forecasts are designed for shorter lead times than those of HRES, ENS, Arpège and PEARP as used in Chapter 5. How does the predicting performance of AROME compare with other weather forecasts? Does using AROME forecasts as additional members improve the overall predicting performance? We recall that HRES is the ECMWF deterministic forecast, ENS is the ECMWF ensemble of forecasts, Arpège is a deterministic forecast from Météo France at a coarser spatial resolution than AROME, and PEARP is the ensemble of forecasts of Météo France based on Arpège.

For this study, we chose 10 power plants from the 219 power plants introduced in Chapter 5. The 10 sites were chosen to test the methods on plants with large capacities and to ensure a geographical coverage of metropolitan France. Compared to the study of Chapter 5, only little changes are to be noticed. The study period still runs from January 2012 to October 2013. Data are still shown after normalization by the plant installed capacity and are consequently dimensionless. Climatological estimations are still generated from a rolling period of 2 months. The statistical models providing conversion between weather forecasts and production forecasts are again trained during the period January 2012 to February 2013, and the remaining time of 2013 is our test

period. AROME forecasts of the 00 UTC analysis are used. We work at the 30-min temporal resolution with clear-sky interpolation of the solar forecasts. Statistical models are built for each lead time (11:00, 11:30, 12:00, ...), which coincides with the hour of the day in this chapter. The main differences with the previous case study are the limited number of 10 sites, the limited lead time of 24 h, and the larger amount of half-hourly models (not only 06:00, 09:00, 12:00, 15:00 and 18:00).

In the following, we introduce how we generate several forecasts thanks to AROME high-resolution. These forecasts are then combined with ML-Poly and CRPS minimization, which is the same online learning technique than previously.

6.1 Building an ensembles of forecasts from AROME forecasts

PV models are trained using almost the same model as in Section 5.1 with only solar radiation forecasts of AROME as input. A reference model is built with the following features: (i) clear-sky normalization of PV power and solar radiation, (ii) smoothing of solar radiation within a $100 \times 100\,\mathrm{km}^2$ square, (iii) linear regression between normalized PV production and solar radiation augmented with non-linear transformations such as square root and square, and (iv) multiplicative bias reduction. The main difference with Section 5.1 is that our reference model for AROME does not include seasonal bias reduction, which failed to provide satisfactory results with AROME.

6.1.1 Leveraging the high spatio-temporal resolution

Spatial smoothing. First, we compare our reference forecasts against models using the almost same features but smoothing areas of $50 \times 50\,\mathrm{km}^2$ and $25 \times 25\,\mathrm{km}^2$. The predictive performance of the 3 forecasts are compared against each other in Figure 6.2 for all sites and all lead time. The MAEs * of the reference forecast beats the $50 \times 50\,\mathrm{km}^2$ forecast by less than 2%, and the $25 \times 25\,\mathrm{km}^2$ forecast by more than 4%. Consequently, the smoothed area of $25 \times 25\,\mathrm{km}^2$ is too narrow for AROME to provide accurate forecasts. We check in Figure 6.3 the difference between the forecasts time series for three consecutive days with high production variability. Graphically we see that all 3 forecasts are quite close from one another. Besides, for homogeneous forecasts during clear sky or very cloudy days, the 3 spatial averages are almost identical. Strategies are described below to obtain a wider variety in the forecasts.

Nearby data in time and space. Since AROME may forecast the presence of a cloud, but not at the correct time or location, an interesting way to leverage this spatio-temporal information is to somehow translate it. For instance, we use the (normalized) solar radiation forecasts for 11:30 as inputs in the 12:00 model to provide additional forecasts for 12:00. We emphasize that this is not equivalent to simply using 11:30 power forecasts for 12:00. Thrice more forecasts than before are hence available thanks to shifts of respectively +30 min and −30 min. Figure 6.4 shows the exact time and

∗. MAE = Mean absolute error of the forecasts $\widehat{y}_t$ compared to the observations y_t, MAE = $\frac{1}{T}\sum_{t=1}^{T} |\widehat{y}_t - y_t|$.

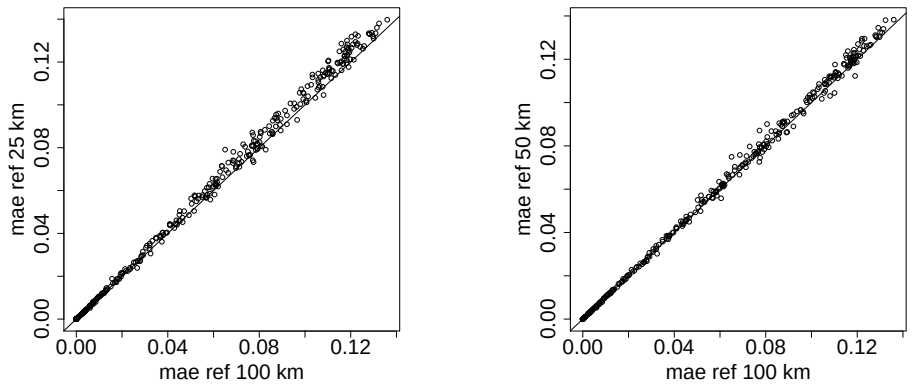


Figure 6.2 – Comparisons of the predictive performance in MAE of the 100×100, 50×50 and $25 \times 25\,\mathrm{km}^2$ models.

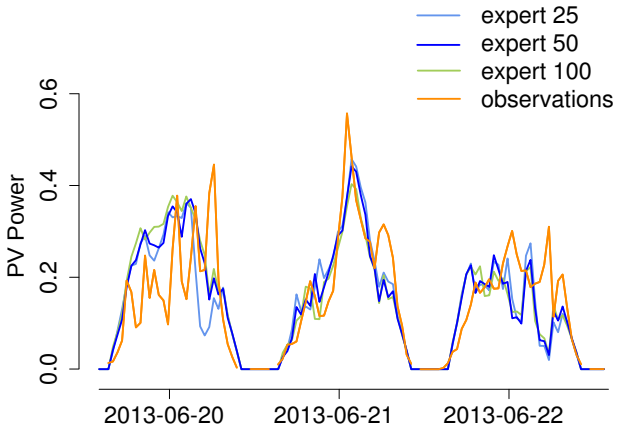


Figure 6.3 – Time series of the forecasts with smoothing areas of 100×100, 50×50 and $25 \times 25\,\mathrm{km}^2$ for one power plant.

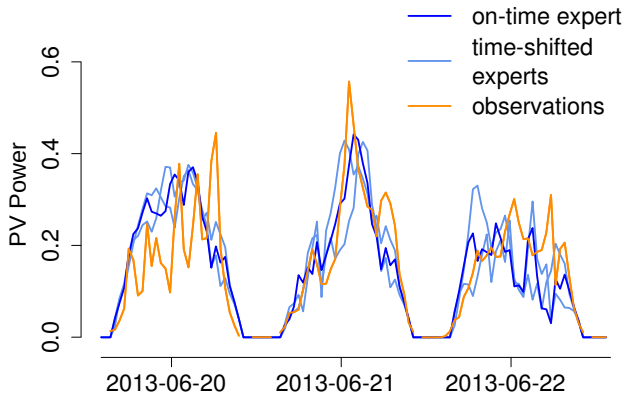


Figure 6.4 – Nearby experts using +30 min and −30 min solar forecasts.

+30 min and −30 min forecasts for several days. We see that the time-shifted forecasts may make up for inaccurate forecasts of ramp events. This is especially true for the second day in Figure 6.4. Yet, the mean production level and the amplitude of the production variations are not correctly described at this point.

In a similar manner, space-shifted forecasts are built using spatial information. Besides the spatial average of the region of averaging, the 0.25, 0.50 and 0.75 spatial quantiles of the region are also used as inputs of the statistical model built with the average solar radiation. In other words for the $100 \times 100\,\mathrm{km}^2$ region, we sort the $40 \times 40 = 1600$ solar radiation forecasts of each 2.5 km. Then the median, the 0.25-quantile and the 0.75-quantile solar radiation values are picked to generate three additional production forecasts. Another option would be to generate forecasts for all the grid-points of the region, but this option is obviously too costly. We show for several consecutive days the production forecasts corresponding to spatial quantiles in Figure 6.5. The mean forecast is flanked by the spatial quantile forecasts. We see that this method may generate forecast peaks, which illustrate the sharp spatial variations of AROME solar radiation.

Model parameterization. Our ensemble of forecasts already comprises 3 (smoothing areas) × 3 (time-shifted) × 4 (spatial estimates) = 36 forecasts, and yet only 3 statistical models were built with different smoothing radii. We kept the 3 smoothing areas because the temporal and the spatial variations seen over the 3 areas differ with the area extension. Depending on the power plant, forecasts focusing on either small-scale or large-scale variations may be of higher interest. The following configurations are also generated:

— Statistical model without clear sky normalization.

— Seasonal bias reduction, described in Section 5.1.

— Clear-sky model with timeshift-optimistic calibration. After clear-sky normalization, we pick the normalized solar radiation in the 3 time-shifted estimates being the closest to the normalized production to train the model. The motivation behind is that such statistical model is presumably built with an artificially good training data set, less sensitive to weather forecasts inaccuracies.

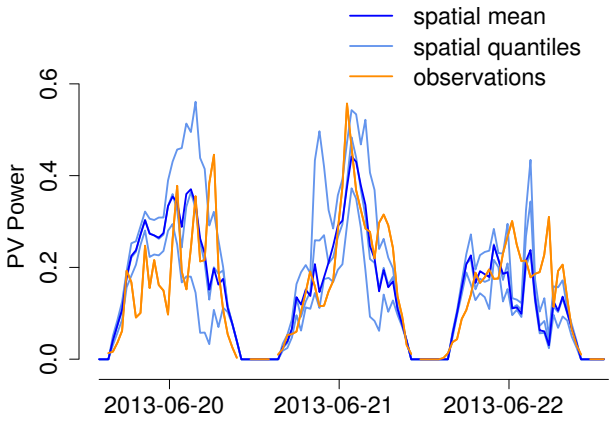


Figure 6.5 – Nearby experts using the 0.25, 0.50 and 0.75 spatial quantiles of the region.

— Cross validation with training blocks of 5 days spaced with buffer blocks of 3 days for a total of 8 training subsets.

A total amount of 336 forecasts are generated with all possible configurations described above.

6.1.2 First sequential aggregation results with AROME meteorological experts

For each lead time, the 336 forecasts are combined with ML-Poly in a similar way as in Section 5.3.

A graphical check with time series. The raw ensemble (with uniform weights) and the weighted ensemble are shown with prediction intervals in Figure 6.6 for France production and also for one of the 10 sites. Four typical days are selected to illustrate a wide variety of situations. The algorithm ML-Poly seems to handle quite nicely the large amount of members. We note that France production appears much smoother than the power plant production, even though France production is here the sum of only 10 sites. For the single power plant production forecasts, strong variations may appear from one time step to another, even though time-shifted experts reduce the amplitude of these variations. Besides, the weighted ensemble seems at least to provide a better estimation of the median of the PDF than the raw ensemble. Both raw and weighted ensembles appear to be under-dispersed for these typical days.

Scores of the combination of AROME forecasts. The bias, MAE, RMSE and CRPS of several forecasts are summarized in Table 6.1 for France production. Once again, the bias of our weighted forecast is quite small, below 0.005. Compared to the raw ensemble (with uniform weights), we show score improvements of 11% for the MAE and the RMSE and 12% for the CRPS. Only little gain is achieved thanks to online learning for MAE and RMSE against AROME reference forecast (1% for the MAE and 3% for the RMSE). We emphasize that our method proceeds to CRPS minimization to generate probabilistic forecasts and not MAE or RMSE minimization for point-

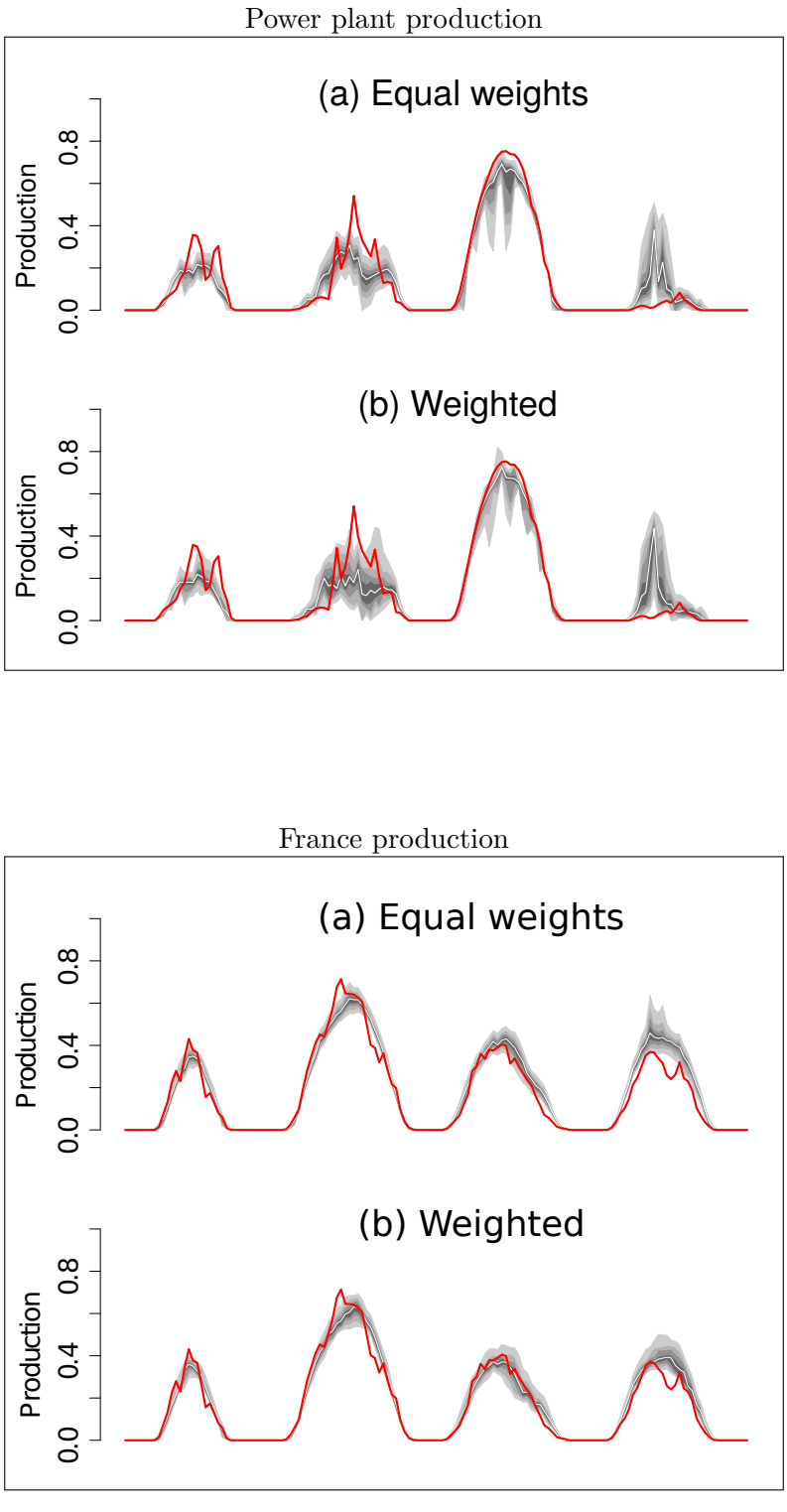


Figure 6.6 – Four typical days of AROME probabilistic forecasts (shaded gray): raw (a) and weighted (b) forecasts for one power plant (above) and for France production (below). Production observations are shown in red.

	Pond	Raw	Climato	AROME Ref
Bias	0.002	0.013	−0.0	0.0
MAE	0.048	0.054 (−11%)	0.089 (−83%)	0.049 (−1%)
RMSE	0.062	0.069 (−11%)	0.110 (−76%)	0.064 (−3%)
CRPS	0.036	0.040 (−12%)	0.061 (−70%)	

Table 6.1 – Forecasts daily scores (bias, MAE, RMSE and CRPS) of France production for our weighted forecast (Pond), the raw ensemble (Raw), the climatological forecast (Climato) and AROME reference forecast (AROME Ref). Score skills against our weighted forecast for France PV power are indicated in parentheses. We emphasize that a skill of -10% means that the score is 10% worst than the corresponding score of our weighted forecast.

forecasts. Similar results are more finely observed with the bias, MAE and CRPS of each lead time for a single power plant and France production in Figures 6.7, 6.8, and 6.9. We clearly see that AROME reference forecast and our weighted forecast are amongst the best forecasts available. Interestingly, our weighted forecast provides larger improvements against the raw ensemble around noon when the production and level of errors are at their highest level.

Other probabilistic verification tools are shown below after additional members are included in the ensemble.

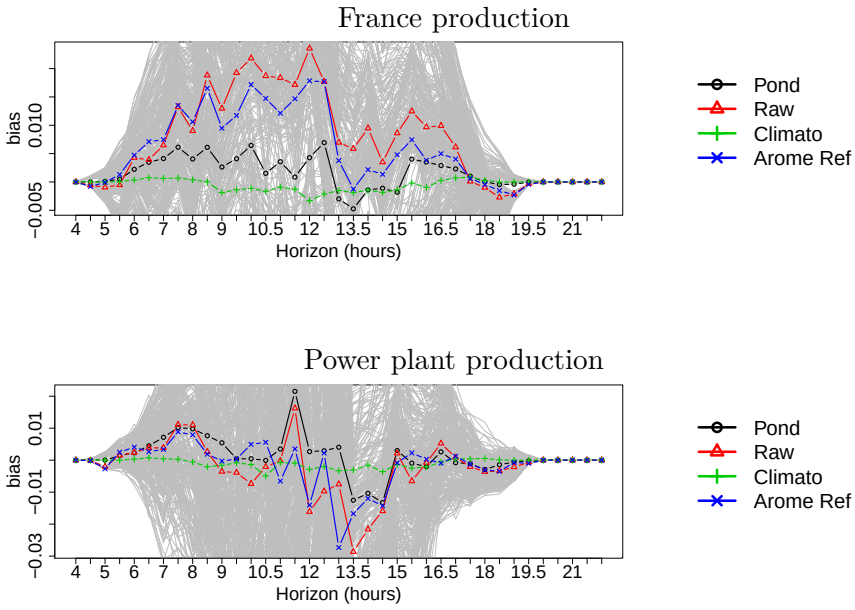


Figure 6.7 – Bias for each half-hour of our weighted forecast (Pond), the forecast with uniform weights (Raw), the climatological forecast (Climato) and AROME reference forecast (AROME Ref) for one power plant and for France production. All 336 forecasts scores are shown in gray.

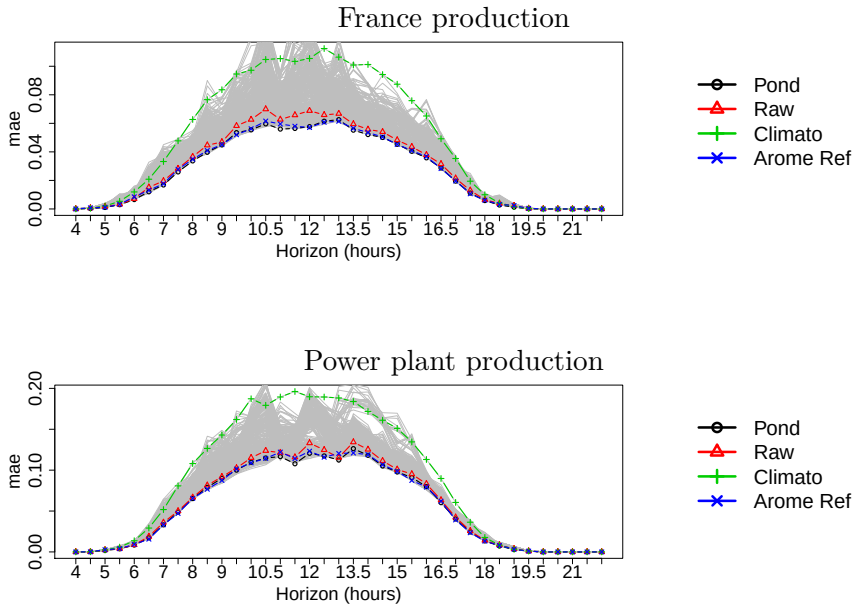


Figure 6.8 – MAE for each half-hour of our weighted forecast (Pond), the forecast with uniform weights (Raw), the climatological forecast (Climato) and AROME reference forecast (AROME Ref) for one power plant and for France production. All 336 forecasts scores are shown in gray.

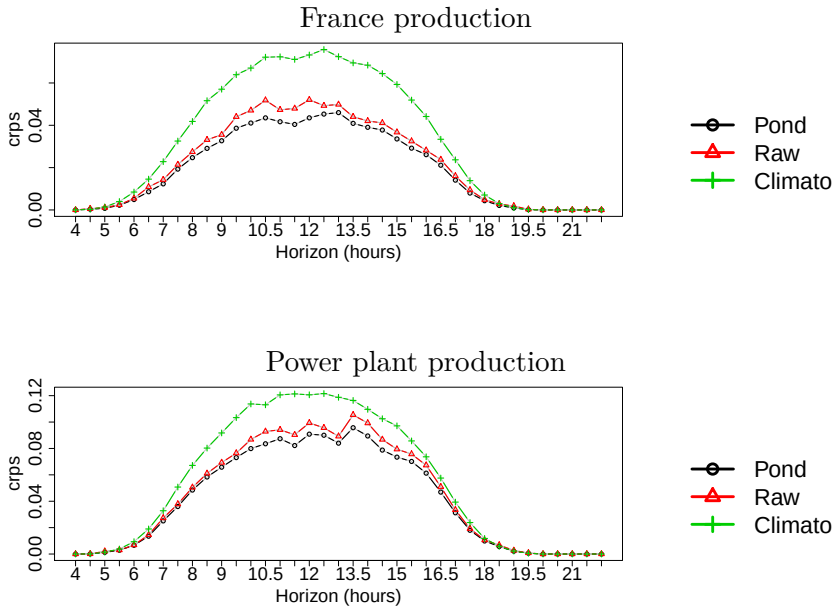


Figure 6.9 – CRPS for each half-hour of our weighted forecast (Pond), the forecast with uniform weights (Raw), the climatological forecast (Climato) for one power plant and for France production.

6.1.3 Adding rolling quantiles experts

Until this point, the experts of our ensemble of forecasts are mainly built using spatio-temporal information of AROME and several parameterizations of our statistical model. The under-dispersion of the raw and the weighted ensemble encourages us to improve our set of experts, by including new experts. This is why we include 2×10 quantile experts in this section using quantile regression (10 experts) and quantile random forests (10 experts). These new experts are learned with rolling training periods of 90 days in order to take advantage of the recent available information. We recall that the previous models are static, trained on 13 months. We target the quantiles $\alpha \in \{0.05, 0.15, \ldots, 0.95\}$ which are optimal for ensemble CRPS minimization [Brö12].

The new quantile experts are learned by using the best former experts as input variables. The former experts are ranked according to their average weight given by ML-Poly during a period of 20 days in January 2013, corresponding to the end of the training period. This criterion is chosen to define the added value of the expert for the ensemble, and to realize a model selection. The 5 experts with highest weights are the input variables to train quantile experts. The motivation behind is that we want to improve upon the already good experts, instead of going again through the statistical modeling process with the meteorological input variables. Also, the members with a very low weight are assumed not to help the weighted forecast. Unnecessary experts with very low weight are hence removed from the ensemble. This concerns about 40% of the experts whose average weight is below 0.001. The daily CRPS of the raw ensemble is improved by removing the least weighted members (6% gain for France production and 4% gain in average for all sites). Further work may investigate the use of online

learning algorithms to build additional quantile experts as in Gaillard et al. [GGN16].

Lasso-penalized quantile regression. Ten new experts are trained with Lasso-penalized quantile regression, inducing sparsity in the model thanks to the L1-norm regularization. In practice, a quantile expert of level α is obtained as a combination $(\boldsymbol{a}_t^\alpha)^\top \boldsymbol{x}_t^{\text{best}}$ of the best members selected above and nonlinear transformations of these experts (square, square root and inverse-logit). The parameters $\boldsymbol{a}_t^\alpha$ are found through minimization of a cost function:

$$\boldsymbol{a}_{d,h}^\alpha = \arg\min_{\boldsymbol{b}} \left[\sum_{d'=d-90}^{d-1} \text{QS}_\alpha(\boldsymbol{b}^\top \boldsymbol{x}_{d',h}^{\text{best}}, y_{d',h}) + \lambda ||\boldsymbol{b}||_1 \right] . \tag{6.1}$$

The notation $\text{QS}_\alpha(x, y)$ refers to the quantile score of level α between the prediction x and the observation y, see Equation 5.4. We adopt here the notation "(day=d, hour=h)" instead of the regular "time=t" to emphasize that quantile experts are learned independently for each half-hour of the day. The parameter λ of the regularization is taken with the default value 0.50.

Quantile random forest. A major difference between random forests and quantile regression is that regression trees resort exclusively to analogue searches, and do not estimate linear or nonlinear effects of the input variables. Random forests [Bre01] introduce randomization compared to regular binary regression trees. A large amount of regression trees are built using random training subsets (500 for example). Besides, only a random subset of the input variables are used at each node tree to find the splitting rule. Given new inputs, the output is commonly the average over the trees of the average value of the observations attached to the final leaf of each tree. Quantile random forest [Mei06] enables to provide estimation of quantiles and not only the mean. To do so, the past observations of the final leaves are gathered to construct a CDF, from which a quantile is deduced.

6.2 Sequential aggregation results with AROME statistically calibrated experts

6.2.1 Improvements with rolling quantile experts

We run the same experiment as described in Section 6.1.2, but with changes in the set of experts. The previous weighted forecast is referred to as Pond.0, while the new weighted forecast is referred to as Pond.1. The new ensemble comprises 3 subsets: the remaining forecasts from AROME, quantile regression experts and quantile random forest.

Half-hourly scores. The CRPS of each half-hour of the new ensemble and its subsets are shown in Figure 6.10. We see that Pond.1 gives better scores than its subsets, and quantile regression experts constitute the second best expert subset.

Average CRPS skill scores. The average CRPS skills against Pond.1 are summarized in Figure 6.11. All subsets of the ensemble taken with uniform weights are beaten by the weighted forecast Pond.1, by 5% for the quantile regression experts, 7% for the raw ensemble, and above 11% for the quantile random forest experts and remaining AROME experts. Besides, the score is improved by 7% compared to the previous

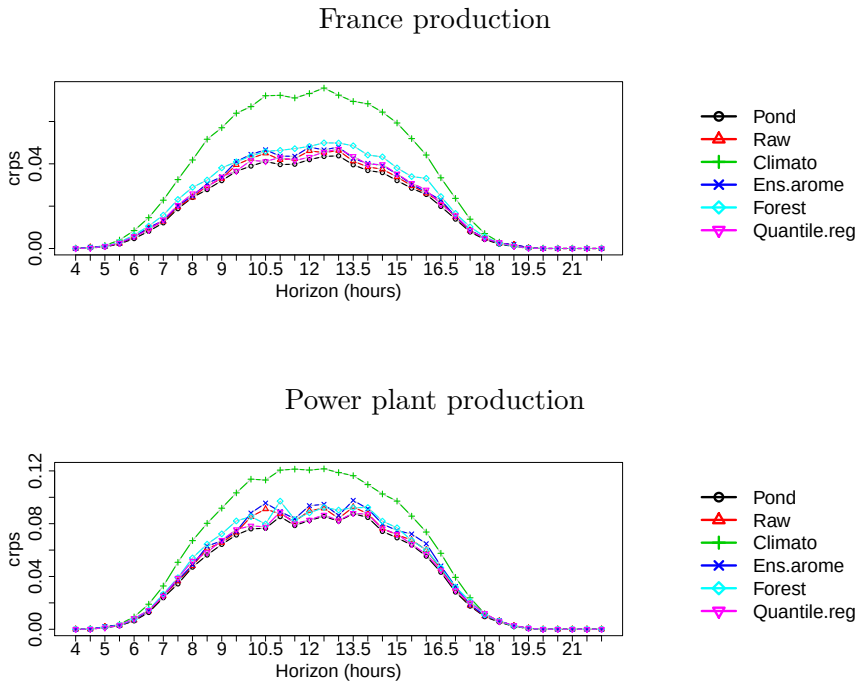


Figure 6.10 – For the ensemble with rolling quantile experts, CRPS for each half-hour of our weighted forecast (Pond.1), the forecast with uniform weights (Raw.1), the climatological forecast (Climato), selected AROME forecasts (Ens.arome.1) quantile regression experts (Quantile.Reg), and quantile random forest experts (Forest) for one power plant and for France production.

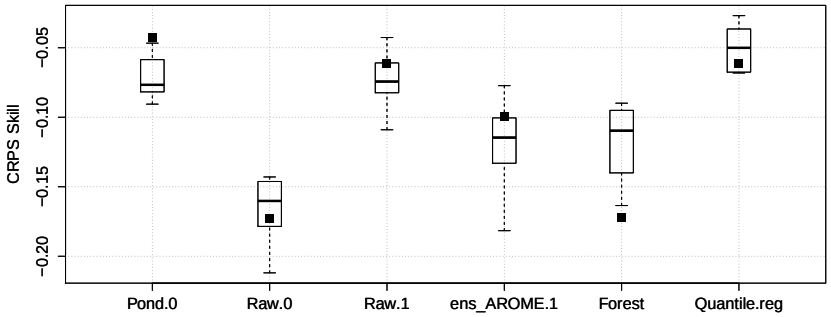


Figure 6.11 – Daily CRPS skill against our new weighted forecast: previous weighted forecast with 336 members (Pond.0) and corresponding raw ensemble (Raw.0), current raw ensemble with added quantile members and removed unnecessary experts (Raw.1), remaining AROME experts (ens_AROME.1), quantile random forest experts (Forest) and quantile regression experts (Quantile.reg). Black squares indicate scores for France production and the score variability is computed with the 10 power plants.

weighted forecast without quantile experts Pond.0. The poor performance of quantile random forest experts, especially for France production, may be due to an insufficient amount of learning data (90 points). These scores and the weight distribution clearly point out that the 20 quantile experts help the weighted forecast Pond.1. Indeed, the weights given to the new experts are more than twice higher than the uniform weight $1/M$ for both quantile regression and quantile random forest experts.

Comparison of Pond.0 with Pond.1. Half-hourly skills of Pond.0 against Pond.1 indicate that our new setting does not improve the MAE, but slightly improves the RMSE around 2%, and CRPS gains are much higher, up to 10%, see Figure 6.12. Hence a major difference between Pond.0 and Pond.1 is the improvement of the spread of the probabilistic forecasts, since the mean of the probabilistic forecast is only slightly improved. A graphical verification of time series supports this statement. We clearly see a spread correction between Pond.1 and Pond.0 in Figure 6.13. Now we check the calibration of Pond.0 and Pond.1. The new weighted forecast shows a rank histogram closer to a flat rank histogram, see Figure 6.14, mainly by reducing the height of the outer bars. This is consistent with the increase of the forecast spread. Also, spread skill diagrams of Pond.1 show a better agreement with the first diagonal than those of Pond.0, thanks to the wider spread of Pond.1 in Figure 6.15. Reliability diagrams for the event "the production is below the climatological median" of Pond.0 and Pond.1 appear to be quite similar in Figure 6.16. The forecast Pond.0 provides better reliability for predicting events with probability below 0.4, and both forecasts tend to over-predict the occurrence of low production with probability between 0.5 and 0.9.

We also investigated the effect of removing the unnecessary experts. The online learning algorithm was run with all $336 + 20$ experts. Interestingly, this weighted combination gives very close results than the weighted combination where unnecessary experts are removed (not shown). Consequently, the online learning algorithm handles quite nicely a large amount of poor experts.

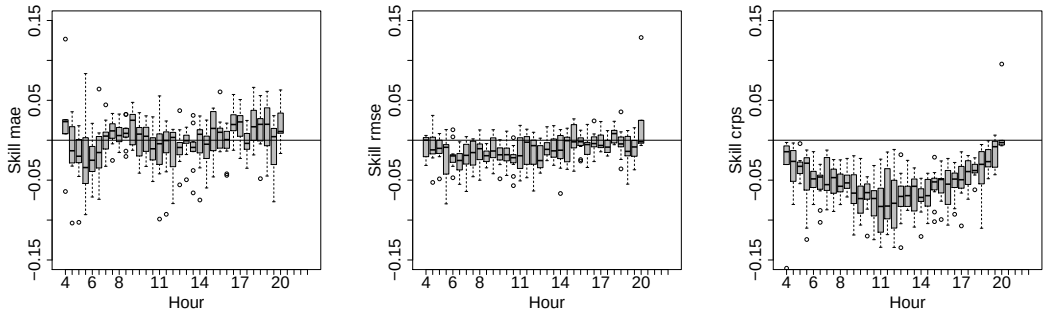


Figure 6.12 – Hourly skill of the weighted forecast with 336 forecasts against the weighted forecast computed rolling quantile experts for the MAE, RMSE and CRPS. Negative skills indicate that better performance are reached thanks to the quantile experts.

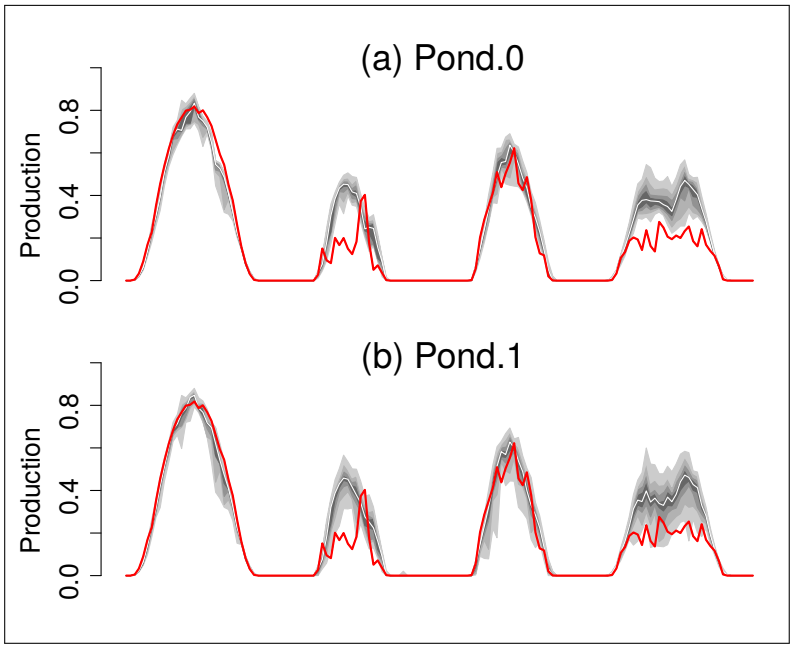


Figure 6.13 – Four typical days of AROME probabilistic forecasts (shaded gray): Pond.0 (a) and Pond.1 (b) forecasts for one power plant. Production data are shown in red.

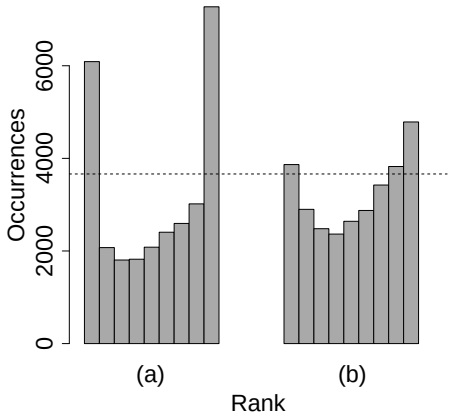


Figure 6.14 – Rank histograms of Pond.0 (a) and Pond.1 (b), computed for the 10 sites and France production for midday hours.

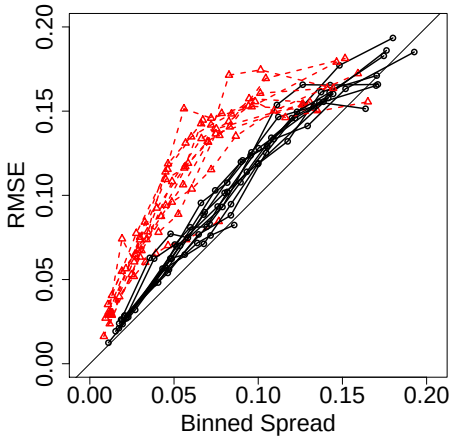
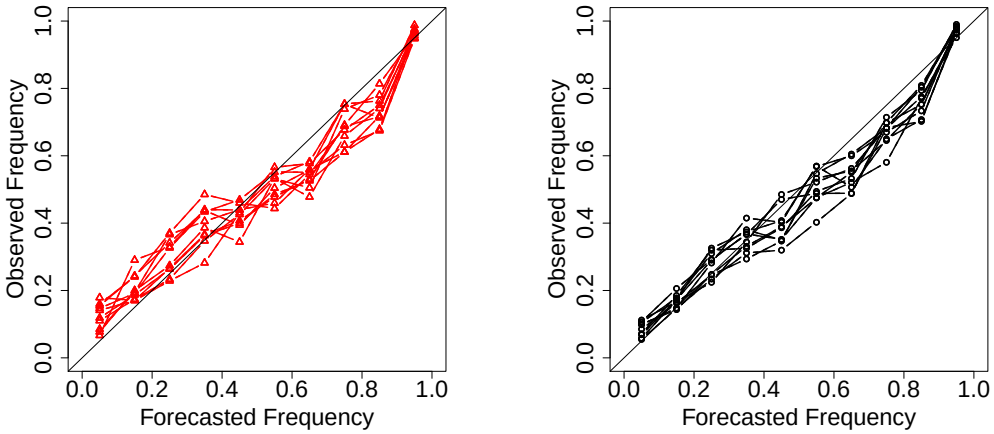


Figure 6.15 – Spread skill of Pond.0 (red) and Pond.1 (black) for each site, computed for the 10 sites and France production for midday hours.



(a) Reliability diagrams of Pond.0.

(b) Reliability diagrams of Pond.1.

Figure 6.16 – Comparison of reliability diagrams between the weighted forecasts, computed for midday hours for each site.

As a conclusion, we used the high-resolution of AROME forecasts to build a large amount of forecasts and deliver probabilistic forecasts. We demonstrated that adding rolling quantile members in the ensemble set improves the quality of the weighted forecasts, mainly for the CRPS thanks to adjustments of the ensemble spread. The online learning algorithm proved its efficiency for the tasks of member selection and improving the calibration of probabilistic forecasts. Further work may investigate the best ensemble of forecasts that one may build with AROME to provide perfectly calibrated forecasts.

6.2.2 Comparison of AROME with other forecasts from Météo France and ECMWF

We now combine the production forecasts introduced in the previous Chapters 5 and 6, from the weather forecasts Arpège, HRES, PEARP, ENS on the one hand and AROME on the other hand. We compare four probabilistic forecasts:

— Pond.smooth derived from Arpège, HRES, PEARP, ENS and quantile forecasts (at a coarser resolution than AROME) in Chapter 5.

— Pond.1 derived from AROME and quantile forecasts in the current Chapter 6.

— Pond.all whose set of experts is the union of the sets of expert of Pond.1 and Pond.all.

— Raw.all the uniformly weighted ensemble of Pond.all.

We emphasize that the algorithm ML-Poly with CRPS gradients is again used to generate the weights of Pond.all. The study is restricted to the 10 PV power plants selected above and the hour 12:00 UTC. Differences in previously shown score are due to the selected hour of verification.

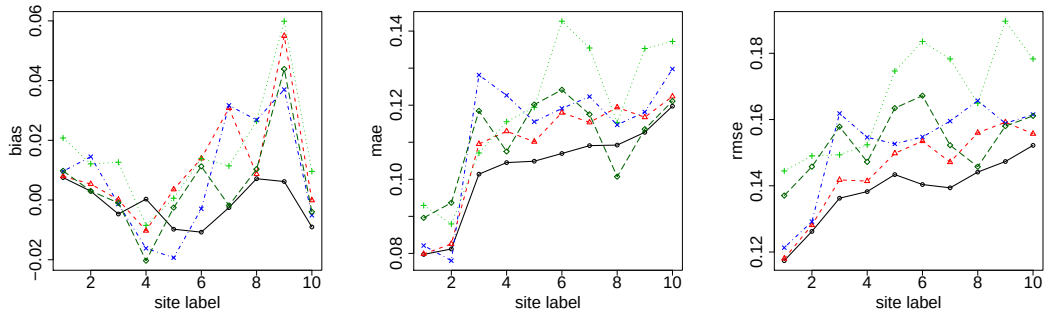


Figure 6.17 – Bias, MAE and RMSE by power plant for the production forecasts from Arpège (green), HRES (dark green), AROME (blue), Raw.all (red) and Pond.all (black). The sites are labeled according to the MAE of Pond.all.

	AROME Ref	Arpège	ECMWF
Skill MAE	-9.8%	-15.5%	-7.5%
Skill RMSE	-9.8%	-20.2%	-10.9%

Table 6.2 – Skill score against Pond.all of the three main deterministic production forecasts.

Comparison of deterministic forecasts. First, we compare the production forecasts derived from AROME, Arpège, HRES, Raw.all and Pond.all in Figure 6.17. We recall that AROME reference forecast is described at the beginning of Section 6.1. The bias of Pond.all is quite low, which is a frequent feature observed with weighted combinations. Besides, for each power plant, the lowest MAE and RMSE is almost always reached by Pond.all. The production forecast from Arpège shows the worst performance in average especially for difficult situations, i.e. for sites with high MAE and RMSE. The sites labeled 1 and 2 correspond to easy situations for Pond.all, especially since AROME forecasts show satisfactory results. The overall weights given to AROME forecasts is indeed higher for these 2 sites than for the other sites (not shown). Score skills against Pond.all are summarized in Table 6.2. The forecast Pond.all shows score improvements of MAE and RMSE above 7.5% compared to HRES, 9% compared to AROME.ref, and 15% compared to Arpège.

Comparison of probabilistic forecasts according to the seasonality and the forecasted level of production. Secondly, we compare the weighted forecasts Pond.1, Pond.smooth and Pond.all. For MAE, RMSE and CRPS, skill scores of Pond.1 and Pond.smooth are worst than -5% against Pond.all, see Table 6.4. The performance of Pond.1 and Pond.smooth vary largely from one power plant to another as shown in Figure 6.18. Indeed, the MAE, RMSE and CRPS of Pond.1 and Pond.smooth show a relative difference higher than 10% for the sites 2 and 4, but Pond.1 beats Pond.smooth at site 2 and Pond.smooth beats Pond.1 at site 4. The scores of Pond.all are almost always lower than those of Pond.1 and Pond.smooth. Hence merging all forecasts improves the weighted forecasts in a wide variety of situations. This statement is supported by the MAE depending on the seasonality and forecasted level of production in Table 6.3. The detailed comparison of Pond.1 and Pond.smooth highlights specific

regimes. During the spring period, we see that it is much harder to forecast correctly for the sites 3-10 than sites 1-2. For the sites 1 and 2, a high level of production is more often forecasted and Pond.smooth is the worst forecast. However, a different regime is observed for the other sites 3-10, where Pond.1 provides better results than Pond.smooth for high level of forecasted production between February and May, but not for low level of forecasted production. Interestingly the converse is observed during summer where Pond.smooth provides much better results than Pond.1 for high level of forecasted production. Although Pond.all is seldom the best forecast with a large margin, Pond.all is very often amongst the best forecasts within a 0.003 MAE margin, which validates its robustness. The only exception concerns the sites 1 and 2 for high level of production during late winter and spring, where Pond.1 clearly beats Pond.all.

The online learning algorithm helps once more to improve the rank histogram of the weighted forecast against the raw forecast, as one can see in Figure 6.19. The added value of the online learning algorithm is also noticeable on the score gains above 4% for the RMSE, 5% for the MAE and 8% for the CRPS, when comparing Pond.all and Raw.all.

Best minimal ensemble subset. To conclude this analysis, we study the best minimal ensemble subset that compares favorably against Pond.all. The online learning experiment was run several times with subset ensembles. We excluded the ensembles PEARP and ENS from this analysis, because we want to show the achievable performance of a weighted forecast using only deterministic forecasts and their related quantile forecasts. The following four subsets were tested {AROME, HRES, ARPEGE}, {AROME, HRES}, {AROME, ARPEGE}, {HRES, ARPEGE}. In other words, adding PEARP and ENS members in the subset {AROME, HRES, ARPEGE} generates the full ensemble of the members that was used for Pond.all, or adding PEARP and ENS members in the subset {HRES, ARPEGE} generates the ensemble used for Pond.smooth. The weighted forecasts from the subsets {AROME, HRES, ARPEGE} and {AROME, HRES} provide very similar results than Pond.all. We found a relative difference inferior to 1% in terms of MAE, RMSE and CRPS between these three weighted forecasts. Consequently, the weather forecasts AROME and HRES seem sufficient to predict as accurately as possible PV power production. Indeed, including information from Arpège, PEARP and ENS does not improve the quality of the forecasts. We note again that adding unnecessary members do not degrade the forecast quality of the weighted ensemble. Other weighted forecasts generated without HRES or without AROME show lower prediction skills, but their performance are still quite close to those of Pond.all. Indeed, we find relative score differences below 8% in terms of MAE, RMSE and CRPS.

6.3 Discussion and perspectives

Probabilistic PV power forecasts are built with the high-resolution weather forecasts AROME. First, multiple forecasts were generated to take advantage of the rich spatio-temporal information of AROME. Secondly, we added rolling quantile forecasts in our ensemble thanks to quantile regression and quantile random forests. These forecasts were combined with online learning techniques so as to minimize the CRPS. We

MAE for sites 1-2				
	Pond.1	Pond.smooth	Pond.all	Number of points
Feb-May level>0.6	0.055	0.058	0.055	68
Feb-May level<0.6	0.087	0.108 $*$	0.090	132
May-Oct level>0.6	0.068	0.067	0.066	203
May-Oct level<0.6	0.117 $\star$	0.150 $*$	0.125	59
MAE for sites 3-10				
	Pond.1	Pond.smooth	Pond.all	Number of points
Feb-May level>0.6	0.086	0.091 $*$	0.085	177
Feb-May level<0.6	0.146 $*$	0.139	0.135 $\star$	583
May-Oct level>0.6	0.079 $*$	0.062	0.064	556
May-Oct level<0.6	0.135	0.143 $*$	0.132	484

Table 6.3 – MAE according to seasonality and forecasted level of production of Pond.all. The asterisk $*$ and the star $\star$ respectively indicate the worst and best score by line. These signs are attributed in case of a clear score difference (above 0.003) .

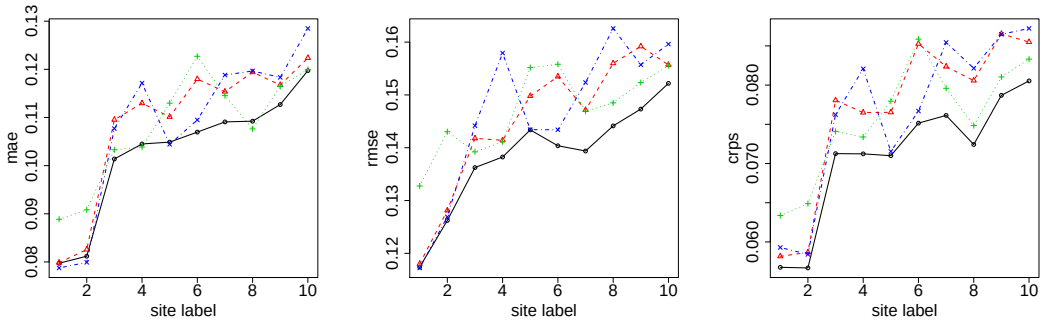


Figure 6.18 – MAE, RMSE and CRPS of the weighted forecasts Pond.1 (blue), Pond.smooth (green), Raw.all (red) and Pond.all (black). The sites are labeled according to the MAE of Pond.all.

	Raw.all	Pond.1	Pond.smooth
Skill MAE	-5.6%	-5.2%	-5.3%
Skill RMSE	-4.7%	-5.7%	-6.2%
Skill CRPS	-8.2%	-7.8%	-6.8%

Table 6.4 – Skill scores of the probabilistic forecasts against Pond.all.

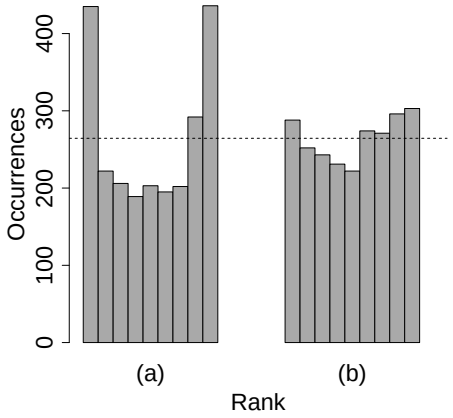


Figure 6.19 – Rank histograms of Raw.all (a) and rank histogram of Pond.all (b).

showed that statistically calibrated members provide a help in the calibration compared to purely meteorological members. Indeed, the rolling quantiles greatly help the calibration of the weighted forecast, even though little improvement is brought to the MAE and the RMSE. We also studied the predictive power of the weighted forecasts including Arpège, AROME, HRES, PEARP and ENS data. We showed that using multiple weather forecasts improves the weighted combination. A minimal subset ensemble of HRES and AROME forecasts proved to deliver accurate forecasts. The online learning algorithm proved to be a serious advantage in the forecast calibration. Indeed, the weighted forecast showed improved calibration compared to its raw ensemble in all cases shown above.

Further work may investigate the calibration of the probabilistic PV forecast with only AROME in order to avoid forecast peaks. Providing an in-depth analysis of the predictive power of the weighted combination using only HRES and AROME data may be of independent interest. An ensemble built with specialized experts based on meteorological regimes could take benefits from the multiple weather forecasts. This raises the questions of which weather forecast and which quantile should be emphasized and under what circumstances. In an other setting, we also could combine solar radiation forecasts first with HelioClim maps or ground data as observation, and then convert the improved solar forecasts to production forecasts. This is left for future research.

7 PV probabilistic forecasts with intraday updates for insular systems

Operational forecasts of Corsica and Réunion include intraday updates of the forecasts with satellite-derived forecasts. These forecasts do not use intraday PV power observations, but only satellite information. We wish to go one step further to improve insular systems PV power forecasts by: (i) generating new predictions with intraday PV power observations, (ii) combining these newly built forecasts with day-to-day predictions and satellite predictions. We show that these new predictions greatly help to improve the quality of the forecasts, and we estimate the accuracy improvement brought by satellite data.

Contents

The insular electric systems of Corsica and Réunion are under the responsability of EDF-SEI. According to the 2016 yearly report on insular electricity production from EDF [*], the install PV capacity in Réunion and Corsica are respectively of 187 MWc and 117 MWc, covering respectively 8.50% and 6.60% of the total production of each island. No interconnection is available for Réunion while Corsica is only connected to Italy. Consequently, the stability of these electric systems may largely suffer from inaccurate PV power forecasts. Réunion and Corsica are two different situations. It is easier to forecast for Corsica, which receives from many clear sky days, while Réunion weather is subject to numerous micro-climates and tropical atmospheric dynamics [LLD16; Bad+15].

In this chapter, we intend to build seamless probabilistic forecasts using operational satellite forecasts and day-to-day forecasts for the total production of each island. Forecasting solar radiation with satellite images is not a new idea [Ham+99], for example tested with HelioClim data [Dam+14]. Besides, combinations of satellite and day-to-day forecasts were found in Lorenz et al. [LKH12], but not for probabilistic forecasting and not with an online learning algorithm. We focus on lead times between 30 min and several hours to bridge the gap between day-to-day forecasts and satellite forecasts, at a 30-min temporal resolution. The operational forecasts may be biased and not correctly calibrated, hence we generate new forecasts using real-time production data and the available forecasts. We emphasize that we rely on a stronger assumption than operational forecasts: the availability of real-time production data. We show that real-time updates make it possible to improve greatly the prediction accuracy compared to day-to-day forecasts. Online CRPS learning is used again to built forecasts combinations. For very short lead times, the challenge is to beat the persistence forecast, which is the projection at time $t - \Delta t$ of the latest available production data $y_{t-\Delta t}$ to time t, by taking into account the clear sky evolution:

$$\widehat{y}^{persistence}(t) = y(t - \Delta t) \times \frac{P_{cc}(t)}{P_{cc}(t - \Delta t)}\,,$$

where $P_{cc}(t)$ is the clear sky production profile for time t.

7.1 Intraday PV updates experimental setup

7.1.1 Operational forecasts

Satellite forecasts of PV power are already operational, using data from the satellite Meteosat 10 for Corsica and Meteosat 7 for Réunion (soon replaced by Meteosat 8). The images are available every 15 min at a 3-km resolution for Corsica for both the visible and infrared spectrums. For Réunion, the resolutions of 2.50 km are available for visible images and 5 km for infrared images every 30 min. We now describe how these images are processed to deliver PV power forecasts. The procedure is applied at over 130 production sites in Réunion and over 20 production sites in Corsica. The total production of each island is the sum of the production of these sites.

*. https://one.edf.gp/sites/default/files/SEI/Panorama%20des%20ENR/panorama_enr_corse_et_outre-mer_2015.pdf

Albedo images from the visible spectrum are transformed into cloud indices after scaling:

$$\text{cloud index} = \frac{\text{albedo} - \text{albedo}_{min}}{\text{albedo}_{max} - \text{albedo}_{min}},$$

to clear the images from information on the ground state and on the hour of the day. The extreme albedo_{min} and albedo_{max} are found with the 3% and 97% quantiles of the 30 previous days. For morning forecasts up to the solar elevation angle of 5°, infrared images of brightness temperature are used. Brightness temperature is the temperature of a black body with the same radiation intensity for this bandwidth. Infrared images are not processed to derive a cloud index.

The forecasting procedure resorts only to image processing and statistical learning but no physics. One statistical model for each site and for each month is trained with spline regression between production data and the pixel of the site in real-time images. In the model, the output production data are normalized by a clear sky profile. We emphasize that the same model with visible images is used for several hours of the day and all lead times. PV power forecasts are generated with these statistical models and cloud index forecasts that we now describe. First, a block-matching algorithm is applied to the last two cloud index images to estimate the cloud motion. Secondly, the cloud motion estimation defines a zone of interest in the last image. The average pixel value of the zone of interest gives the forecasted cloud index. The zone of interest is a triangle with one vertex at the site location and a 20° angle. The triangle direction is the upstream cloud motion and its height is set by the lead time and the cloud motion speed.

We emphasize again that the operational satellite forecasts uses real-time data from satellites but no real-time PV power observations. Production data are not perfectly reliable, because data transmission from production meters to grid operators may fail.

Operational day-to-day production forecasts are also available. They rely on solar radiation, total cloud cover and 2-m temperature HRES forecasts from ECMWF of the 00 UTC analysis. Statistical models are built to deliver a deterministic production forecast using ECMWF data and quantile predictions associated with the deterministic forecast. These statistical models are similar to those of Chapter 5.

In the following, the operational satellite forecast is referred to as “prevsat” and the operational HRES forecast is referred to as the day-to-day forecast or “prevenir”.

7.1.2 Building new forecasts with intraday updates

As in Chapter 6 with AROME forecasts, we generate rolling quantile predictions with Lasso-penalized quantile regression and quantile random forests. These quantile forecasts are trained with 5 major input variables:

— the last production forecast from either “prevenir” or “prevsat”,

— the last production data available,

— their corresponding cumulated value over the current day to indicate the average level of production and forecast for this day,

— the production data of the same hour of the previous day.

Name	Description
Pond	weighted forecast
Raw	raw ensemble with uniform weights
prevenirQ	quantile day-to-day forecasts
qr_prevsat	rolling quantile regression with satellite forecasts
qrforest_prevsat	rolling quantile random forests with satellite forecasts
qr_prevenir	rolling quantile regression with day-to-day forecasts
qrforest_prevenir	rolling quantile random forests with day-to-day forecasts

Table 7.1 – Summary of probabilistic forecasts in the set of experts.

Past day production data is especially interesting to predict very low level of production during sunrise and sunset. The statistical models use normalized data by a clear sky production profile, and the cumulated values are normalized by a cumulated clear sky profile. We generate models for each lead time and each half-hour of the day with a rolling training period of 3 months. For quantile regressions, nonlinear transformations (square, square root and inverse logit) of the 5 major input variables are included in the model to reach the total of 20 input variables.

A total of 60 new forecasts are generated with 15 level of quantiles, 2 statistical methods (quantile regression and random forests) and 2 input variable sets (day-to-day and satellite forecasts). Larger training sets are obtained with available data of nearby half-hour of the day. For instance, the (12:00 UTC, 1-h lead time) model is trained with the data initially gathered for the (11:30 UTC, 1-h lead time), (12:00 UTC, 1-h lead time) and (12:30 UTC, 1-h lead time) models.

In the full ensemble of experts, the following 82 forecasts are present:

— 1 deterministic satellite forecast “prevsat”,

— 1 deterministic day-to-day production forecast “prevenir” and 19 corresponding (non-rolling) quantiles “prevenirQ”,

— 60 rolling quantiles introduced in this section,

— 1 persistence forecast “Persistence”.

The forecasts are summarized in Table 7.1 with notation for each subensemble. Climatological productions (Climato) are also shown in the figures to illustrate the gain brought by forecasts.

7.1.3 Online learning experiment

The data sets range from May 2014 to June 2015 for Corsica and from September 2015 to September 2016 for Réunion. The online learning experiment starts at the beginning of these periods. The test periods begin 3 months later due to first rolling period. The satellite statistical models are trained in 2010-2011 for Réunion and 2013-2015 for Corsica. The day-to-day statistical models are trained in 2012-2014 for Réunion and from 2012 to June 2015 for Corsica. Our experiment is therefore not fully operational since the training and the testing periods of “prevsat” and “prevenir”

overlap for Corsica. We show that the rolling quantile predictions are critical to improve the forecast accuracy. Since the rolling quantile predictions do not have any overlap between training and testing, we think that the conclusions of this chapter are rather robust.

We use the algorithm ML-Poly with CRPS gradients, as in Chapter 6 and 5. The weights were initially updated once per day for each lead time and half-hour of the day. As an example the weight of the mth member with 30-min for the 12:00 UTC PV production of day d that we note here $u_{m,12:00,+30min,d}$ was updated to $u_{m,12:00,+30min,d+1}$ after $y_{12:00,d}$ is received. Because the testing periods last less than a year, we use a trick for the weights to see more data. For a fixed lead time, the available data of nearby half-hour the day also produce a weight update. For instance, the observations and the forecasts of 1-h lead time of 11:30 UTC, 12:00 UTC and 12:30 UTC contribute to the update of the (12:00 UTC, 1-h lead time) weights. This setting is still operational because the weight updates can be achieved at the end of each day. The weighted combination of the forecasts is referred to as the weighted forecast.

7.2 Results

7.2.1 Time-series, spread and weights

Time-series of the weighted forecasts of Réunion and Corsica are shown in Figure 7.1 and 7.2. The day-to-day forecasts are also shown to illustrate their large spread and their smooth aspect. The error and the spread of the weighted forecasts increase with the lead time, hence more accurate forecasts are delivered for short lead times. Production sharp variations affect the weighted forecasts after the delay corresponding to the forecast lead time. The forecasts sharp variations are due to the rolling quantile forecasts and not to the weight updates, which occur at the end of the day.

The evolution of the probabilistic forecasts spreads in Figure 7.3 clearly indicates the level of uncertainty associated to each subensemble. For short lead times, the spread of rolling quantile forecasts is quite small thanks to the production data within the model inputs. For long lead times of a few hours, the spread of satellite-based forecasts are much larger than the spread of day-to-day forecasts, providing hints on the predictive power of satellite-derived forecasts for such long lead times. We also observe that quantile random forests deliver forecasts with larger spread than quantile regression while the same model inputs were used, and that probabilistic forecasts have a larger spread for Réunion than for Corsica. Besides, the raw ensemble shows quite a large spread, compared to those of our weighted forecast.

The average weight given to each member according to the lead time highlights the predictive power of the subensembles, see Figure 7.4 for Réunion and Figure 7.5 for Corsica. The small spread of the weighted ensemble at short lead times is due to the high weights given to forecasts with small spread, i.e. rolling quantile predictions. The weights given to the day-to-day forecasts increase with lead time, while the weights of the persistence forecast decrease with the lead time. This illustrates a trade-off between daily weather information and production information. The forecasts resulting from quantile random forests often receive lower weights for their extreme quantiles. This

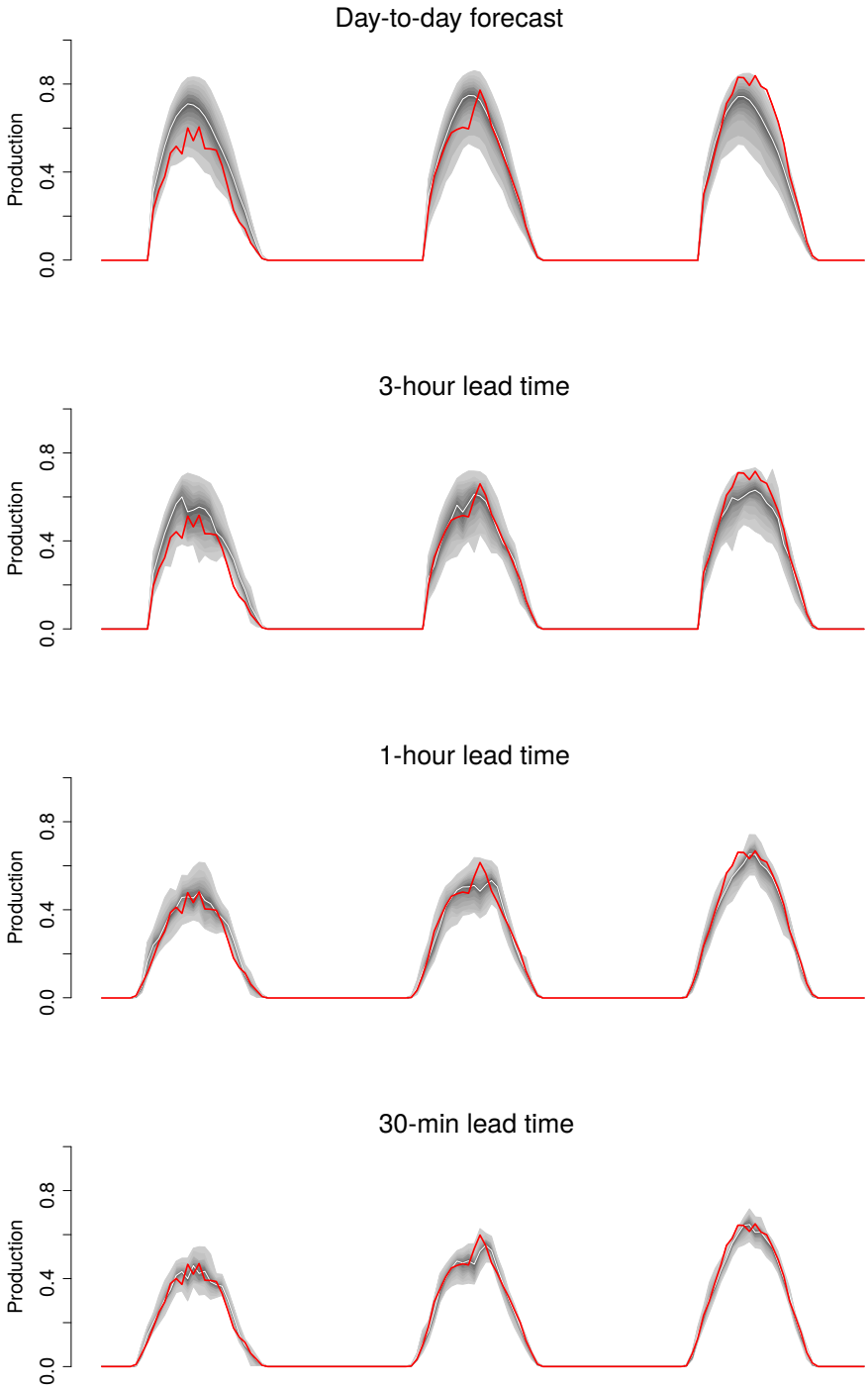


Figure 7.1 – Time series of our weighted probabilistic forecast for 3 consecutive days for Réunion PV production (observation in red). The differences in the time series of intraday lead times of 3 hours, 1 hour and 30 minutes show the evolution of the forecast sharpness and precision. The day-to-day forecast (above) does not benefit from intraday updates.

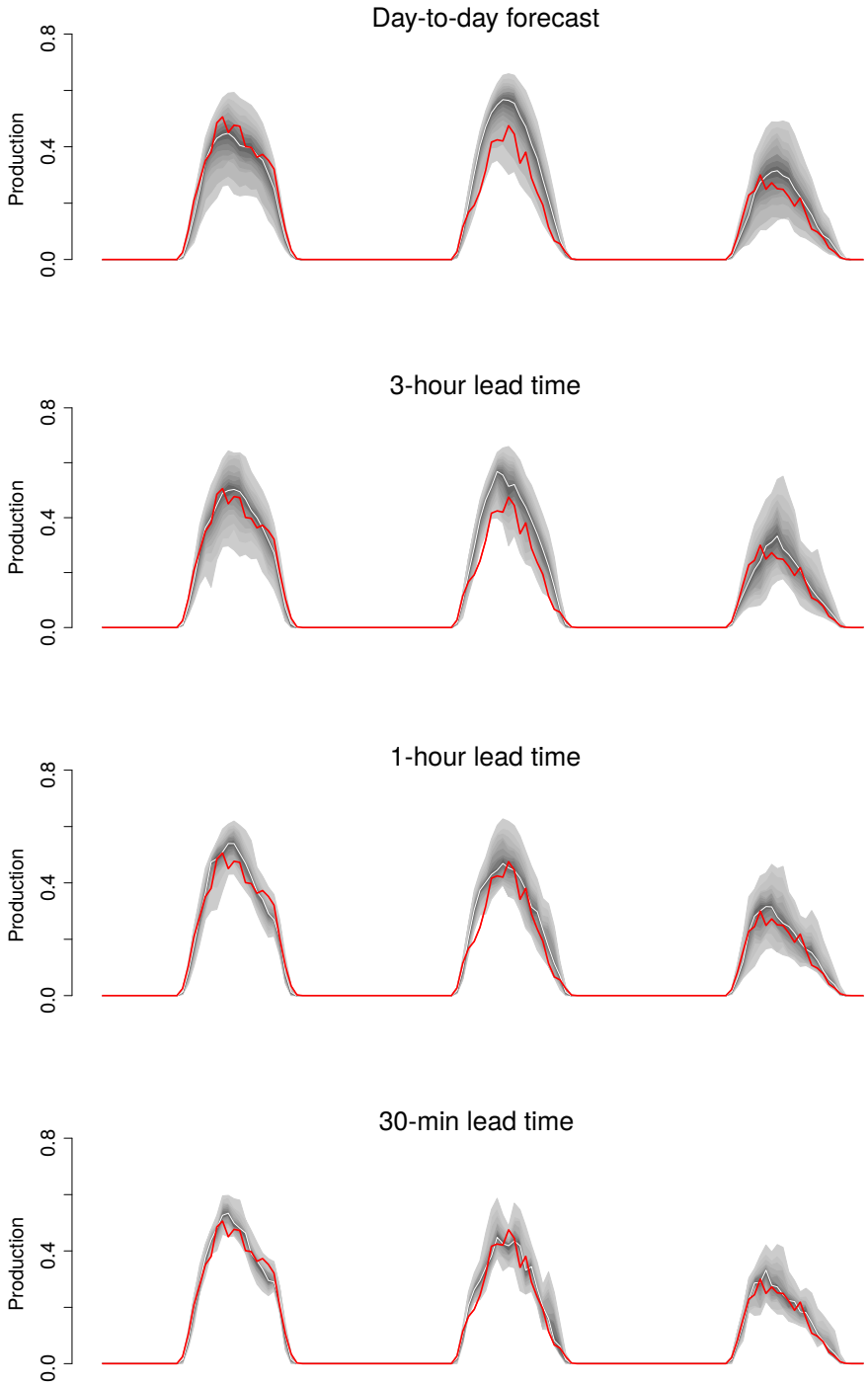


Figure 7.2 – Time series of our weighted probabilistic forecast for 3 consecutive days for Corsica PV production (observation in red). The differences in the time series of intraday lead times of 3 hours, 1 hour and 30 minutes show the evolution of the forecast sharpness and precision. The day-to-day forecast (above) does not benefit from intraday updates.

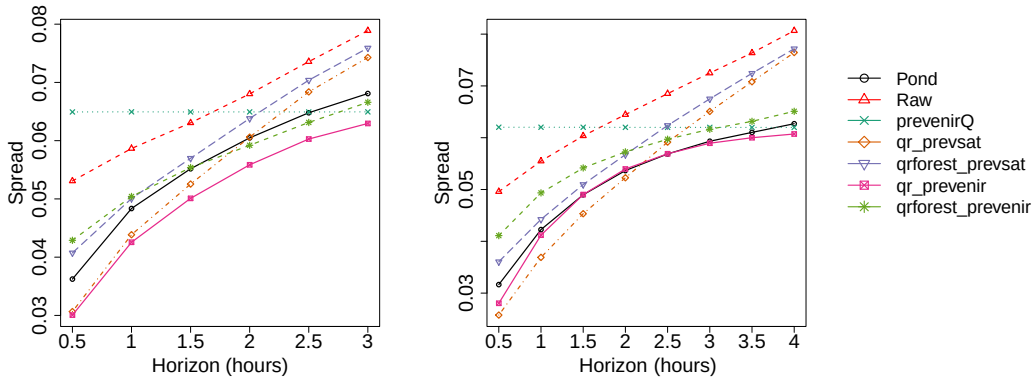


Figure 7.3 – Evolution of the average spread of probabilistic forecasts with the lead time for Réunion (left) and Corsica (right). Morning and evening data are not included in these spread averages.

may be due to the larger spread observed for quantile random forest subensembles. Moreover, satellite-derived forecasts receive higher weights for Corsica than for Réunion. To conclude this analysis, we note that quantile regression forecasts are more prone to receive lower weights than quantile random forest forecasts, while the former exhibit better performance than the latter. We also highlight the difficulty to interpret the average weights of one forecast, because they heavily rely on the behaviour of the other members. For example, 2 median forecasts with similar quality may either share the overall weight that would have been attributed if only one median forecast is present in the ensemble, or one median forecast may prevail over the other one, the latter receiving a very low weight compared to the former.

7.2.2 Probabilistic forecasts performance and calibration

Results for the 30-min **lead time.** We now focus on the scores obtained for the 30-min lead time, when satellite forecasts are possibly at their best, see Figure 7.6 and 7.7. Rolling quantiles, using real-time production, obviously predict much better than the initial forecasts, both satellite and day-to-day forecasts. This is verified for the rolling subensembles with the CRPS and for the rolling median quantiles for the MAE. We note the difficulty to beat the persistence forecast for this short lead time, while it is part of the rolling median inputs. The large MAE of the satellite forecast "prevsat" in the morning at Réunion (worst than climatology) are due to large biases in the forecasts. Moreover, the satellite forecasts at Corsica are not better than day-to-day forecasts for midday hours. The fact that a single model is used for all half-hours of the day may explain the poor performance of "prevsat", and supports our motivation to build models for each half-hour. Besides, the high weights given to satellite-derived quantile forecasts at Corsica encourages us not to drop satellite forecasts, but supports the need to improve the initial forecast "prevsat". Our weighted ensemble is amongst the best forecasts, but is beaten in MAE by the persistence forecast at Réunion.

Score daily averages. Daily scores are average scores over all half-hours weighted

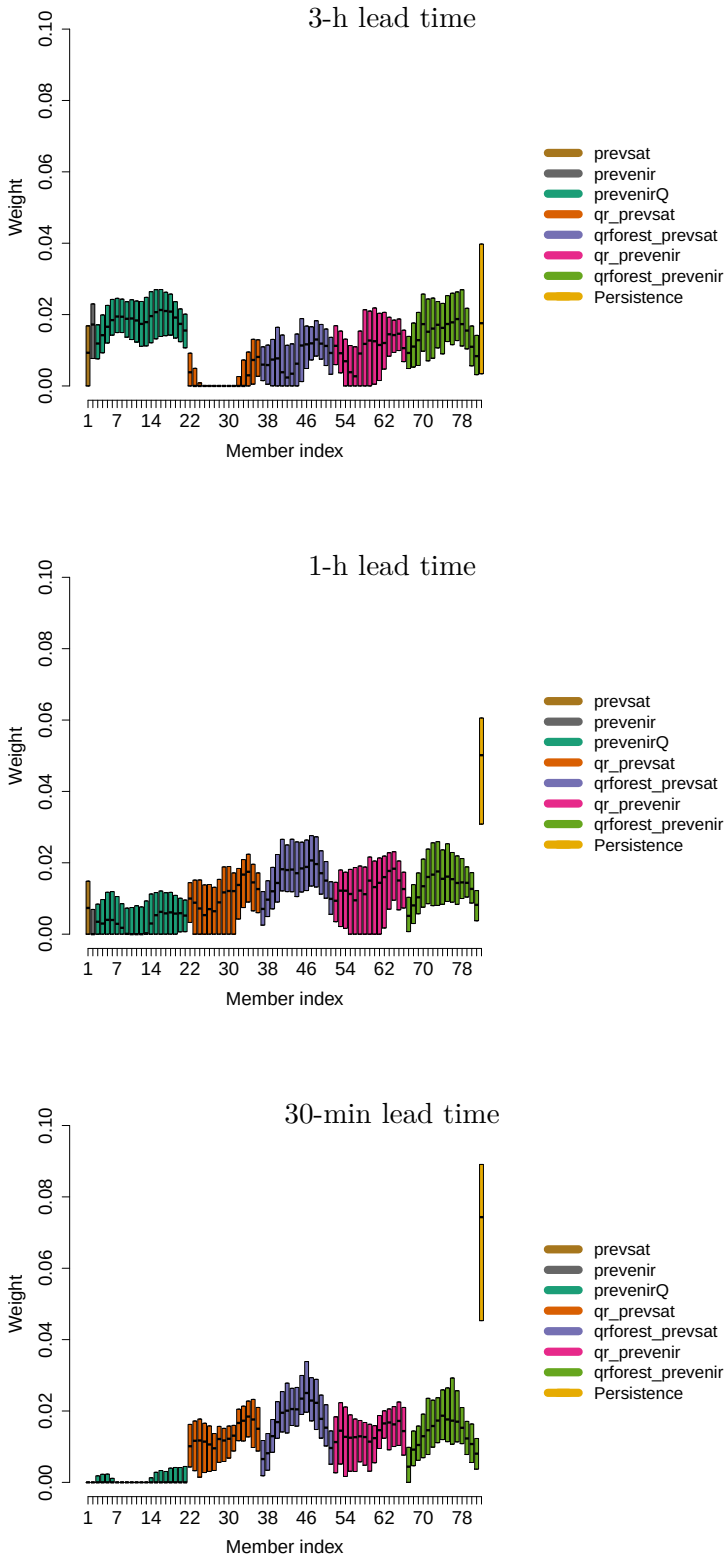


Figure 7.4 – Weight boxplot of each member according to the lead time for Réunion. The boxplot bounds are the first and third quartiles. Morning and evening data are not included in these weight averages.

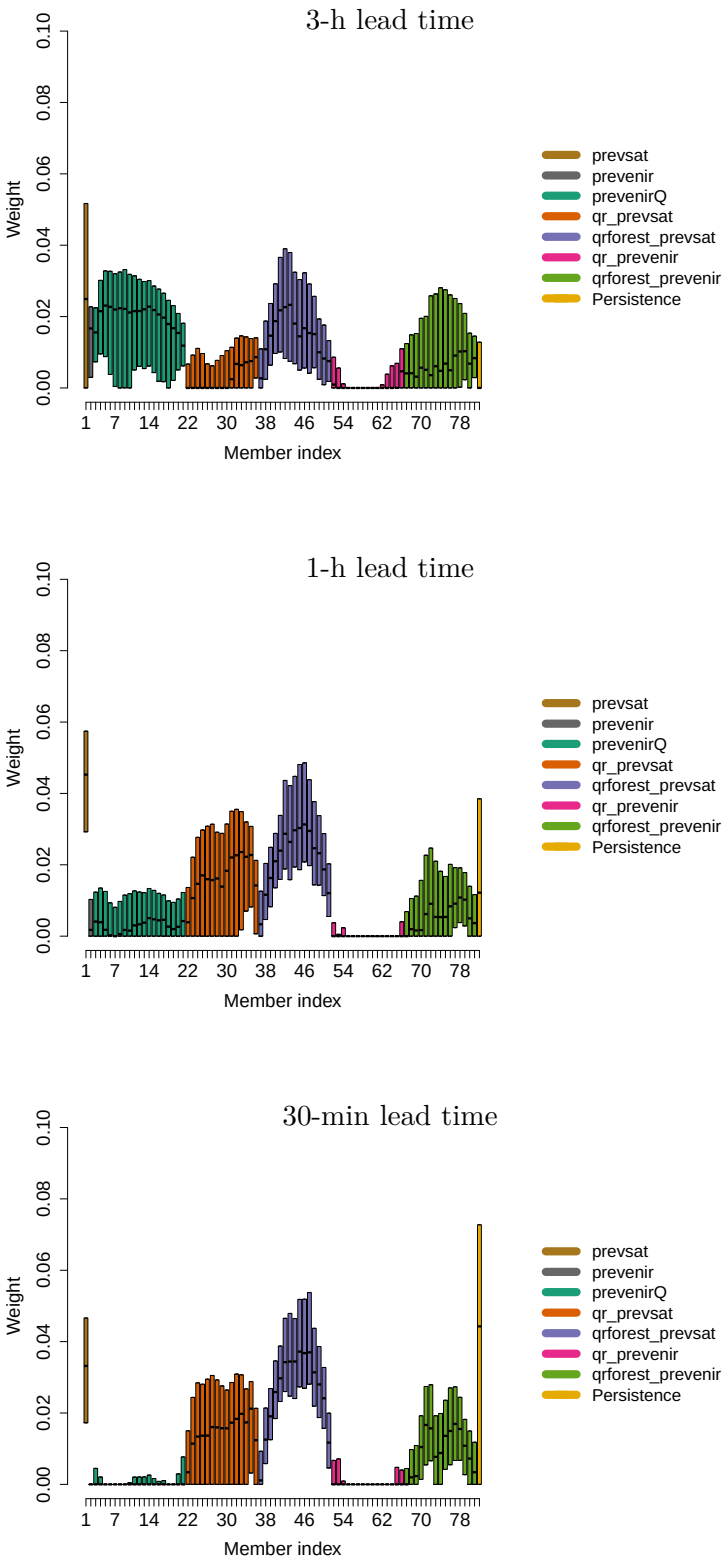


Figure 7.5 – Weight boxplot of each member according to the lead time for Corsica. The boxplot bounds are the first and third quartiles. Morning and evening data are not included in these weight averages.

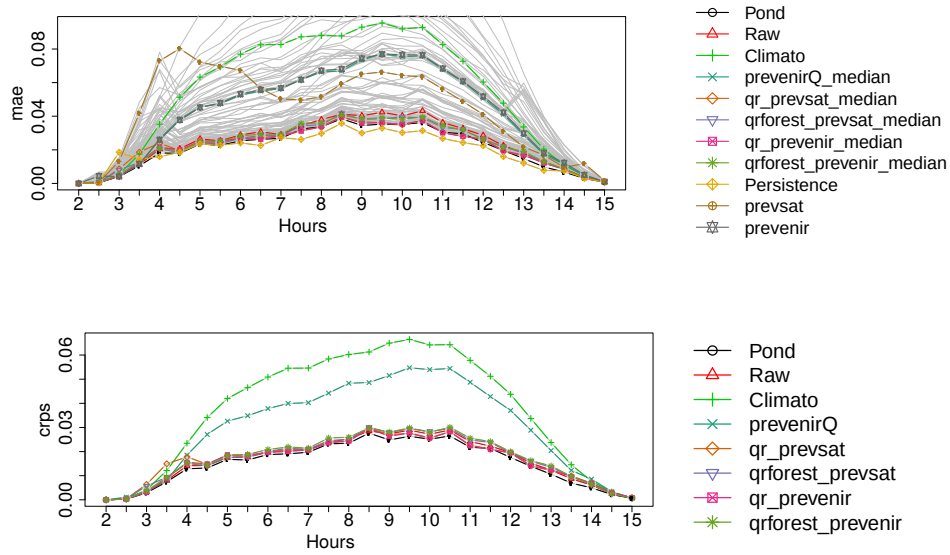


Figure 7.6 – MAE (top) and CRPS (bottom) half-hourly scores for 30-minute lead time at Réunion. The MAE of all individual forecasts are shown in gray.

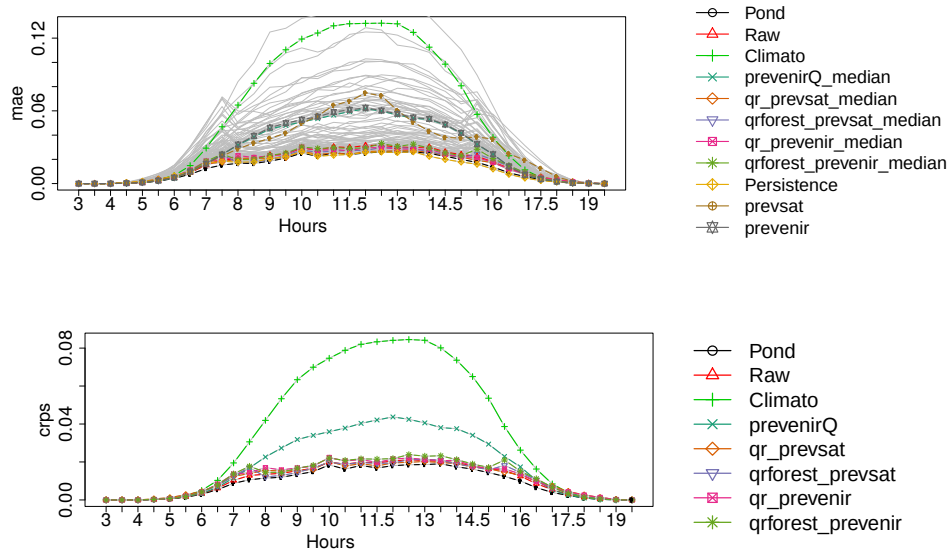


Figure 7.7 – MAE (top) and CRPS (bottom) half-hourly scores for 30-minute lead time at Corsica. The MAE of all individual forecasts are shown in gray.

by the mean observation as introduced in Equation 5.10. Our weighted forecast ranks first for almost all lead times and scores. As expected, the quality of the forecasts gets worse with increasing lead time, see Figures 7.8, 7.9, 7.10 and 7.11. The higher difficulty of forecasting for Réunion is seen with the scores of the weighted forecast. The MAE for Réunion lies between 0.029 (30 min) and 0.057 (3 h), while the MAE for Corsica lies between 0.021 (30 min) and 0.046 (4 h), producing a score relative difference above 20%. The satisfactory performance of the persistence forecast is tempered by its poor quality for lead times longer than 90 min.

The satellite-derived quantile forecasts "qr_prevsat" are better suited for short lead times, especially for Corsica where they largely beat "qr_prevenir" and "qrforest_prevenir" quantile forecasts deriving from day-to-day forecasts (about 10% better). However, they show worst performance than day-to-day forecasts (with no updates) for lead times longer than 2 h in both Réunion and Corsica. In fact, day-to-day forecasts are amongst the best forecasts for long lead times, because little information is added from observations dating from several hours earlier. Interestingly, quantile regressions "qr_prevenir" and "qr_prevsat" provide respectively better results than the corresponding forecasts with quantile random forests, but they receive lower weights than their random forests counterparts.

The raw ensemble is always amongst best forecasts in terms of CRPS, but not for the MAE especially for short lead times. Consequently, including day-to-day forecasts in the ensemble set seems to degrade the distribution mean, but to improve the distribution spread for short lead times. In other words, the raw ensemble takes benefits from the large ensemble spread of day-to-day forecasts for the CRPS but not for the MAE, which depends only on the ensemble mean.

For Corsica at intermediate lead times of 120 min and 180 min, we observe noticeable effects of model blending since the raw ensemble and the weighted forecasts have better scores than other forecasts. These high gains may be due to the high diversity between the individual forecasts. Although it is difficult to show causality, we highlight the correlation between these two facts through the higher difference in Corsica between rolling quantile forecasts from day-to-day and satellite data. We checked the average absolute difference between the median quantiles of "qr_prevenir" and "qr_prevsat", and found that this average absolute difference is twice higher for Corsica than Réunion for the lead time of 30 min and 30% higher for the lead time of 3 h.

Added value of satellite information. In order to assess the utility of day-to-day and satellite forecasts, we run the online learning experiment on two ensemble subsets, either using satellite information or day-to-day forecasts. The persistence forecast is included in each subset. The first subset comprises "prevsat", "qr_prevsat", "qrforest_prevsat" and "Persistence", while the second subset comprises "prevenir", "prevenirQ", "qr_prevenir", "qrforest_prevenir" and "Persistence". The weighted forecast without satellite information is referred to as "Pond.prevenir" and the weighted forecast without day-to-day forecasts is referred to as "Pond.prevsat". Results are shown in Figure 7.12. For Réunion, we find that satellite forecasts do not bring a significant amount of information since very similar forecasts are obtained with and without them. On the contrary, satellite forecasts are quite useful for Corsica. Indeed, they allow a 10% gain up to 2 h of lead time against the weighted forecast "Pond.prevenir" (without satellite

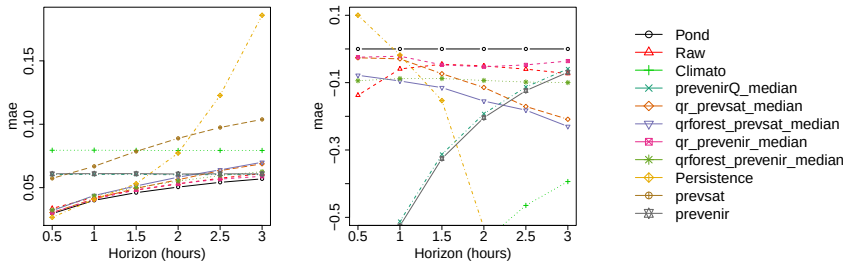


Figure 7.8 – Evolution of the MAE with lead time for all half-hours at Réunion, net values (left), skill scores against the weighted forecast (right).

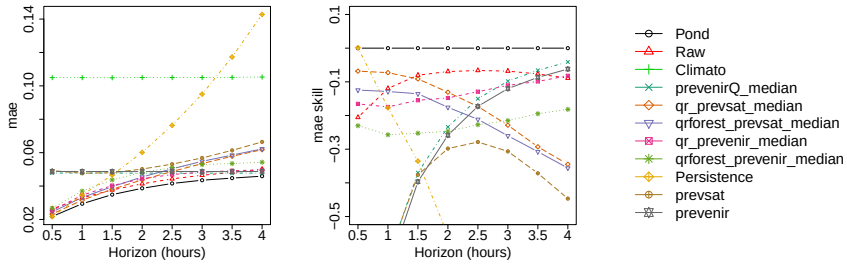


Figure 7.9 – Evolution of the MAE with lead time for all half-hours at Corsica, net values (left), skill scores against the weighted forecast (right).

information). Besides for Corsica, the weighted forecast without satellite information "Pond.prevenir" takes advantage of the persistence forecast for short lead times, which demonstrates once again the importance of real-time PV power observations for short lead times.

Probabilistic forecasts calibration. The calibration of the weighted forecasts and the raw ensemble are now checked to validate the help brought by the online learning algorithm. Rank histograms are shown in Figure 7.13. We see that the weighted forecasts have much flatter rank histograms, compared to the over-dispersed raw ensemble. This statement is especially true for short lead times. Yet, our weighted forecasts show a light under-dispersion, clearly seen at the outer bars. Including a finer description of the rolling quantiles (61 quantiles instead of 15 quantiles) did not improve the weighted

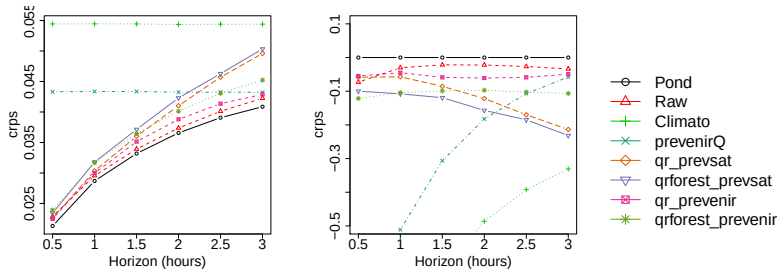


Figure 7.10 – Evolution of the CRPS with lead time for all half-hours at Réunion, net values (left), skill scores against the weighted forecast (right).

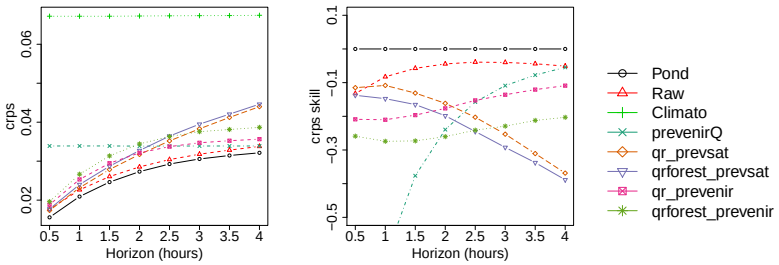


Figure 7.11 – Evolution of the CRPS with lead time for all half-hours at Corsica, net values (left), skill scores against the weighted forecast (right).

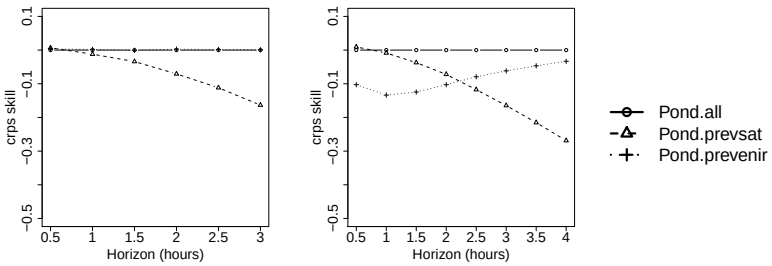


Figure 7.12 – CRPS skills of the weighted forecasts with restricted subsets for Réunion (left) and Corsica (right) against Pond.all (the weighted forecast with all members).

forecast calibration (not shown).

The spread-skill diagrams validate many aforementioned statements from the rank histograms, from the evolution of the spread and from the evolution of the scores with the lead time. We see the strong error increase between the lead times of 30 min and 3 h for several level of errors in Figure 7.14. Here the over-dispersion of the raw ensemble is exposed, especially for short lead times. The light under-dispersion of our weighted forecast concerns small level of errors for Réunion, while an over-dispersion of our weighted forecasts is revealed for large levels of errors at Corsica. The improvements brought by online learning are clearly demonstrated, since the spread-skill diagrams of our weighted forecasts match better the first diagonal. The case of Réunion at long lead times mitigates this statement.

We conclude this analysis with reliability diagrams for the event “production is below the climatological production”. They show satisfactory agreement with the first diagonal, especially for long lead times in Figure 7.15. The questionable reliability of our weighted forecast for short lead times may be due to the tighter spreads of the subensembles.

Conclusion and perspectives

Probabilistic PV power forecasts were built for Réunion and Corsica for lead times of 30 min to several hours. They relied on multiple forecasts based on day-to-day forecasts, satellite forecasts and real-time PV power observations. The multiple forecasts were

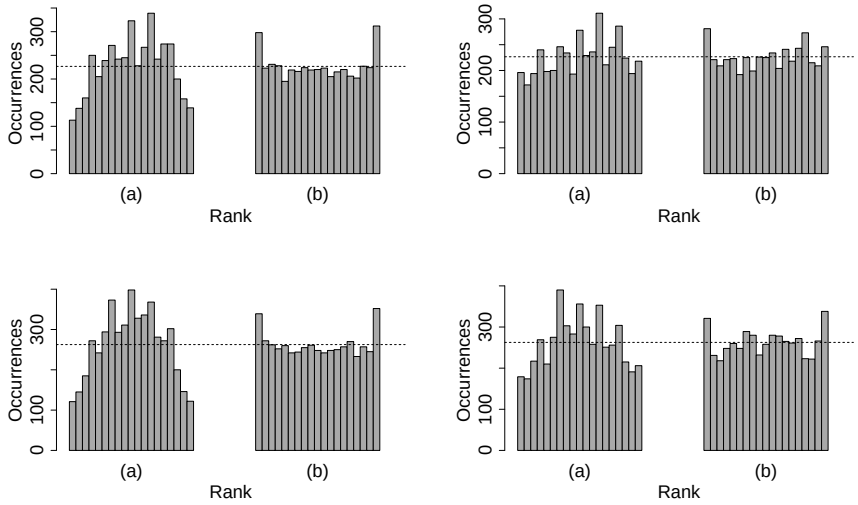


Figure 7.13 – Rank histograms of the raw ensemble (a) and weighted ensemble (b) for midday hours: 30-min lead time Réunion (top left), 3-h lead time Réunion (top right), 30-min lead time Corsica (bottom left), 3-h lead time Corsica (bottom right).

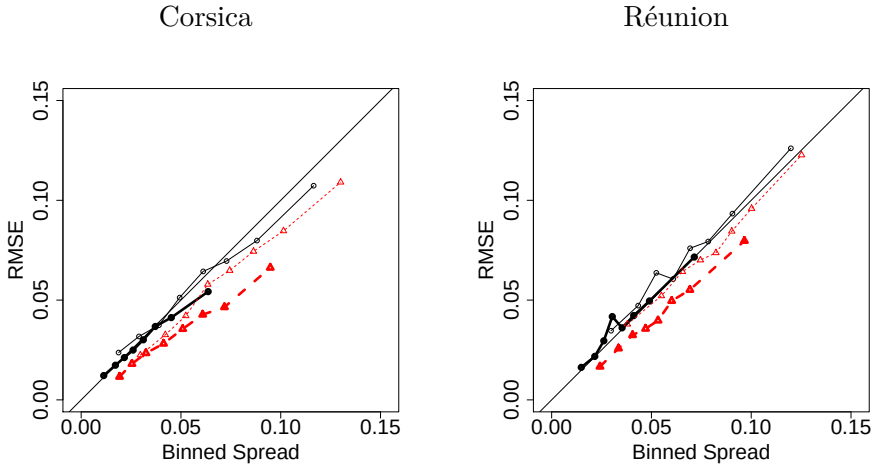


Figure 7.14 – Spread-skill diagram for Corsica and Réunion for the lead times of 30 min (bold line) and 3 h (thin line) for the weighted forecast (black) and the raw ensemble (red).

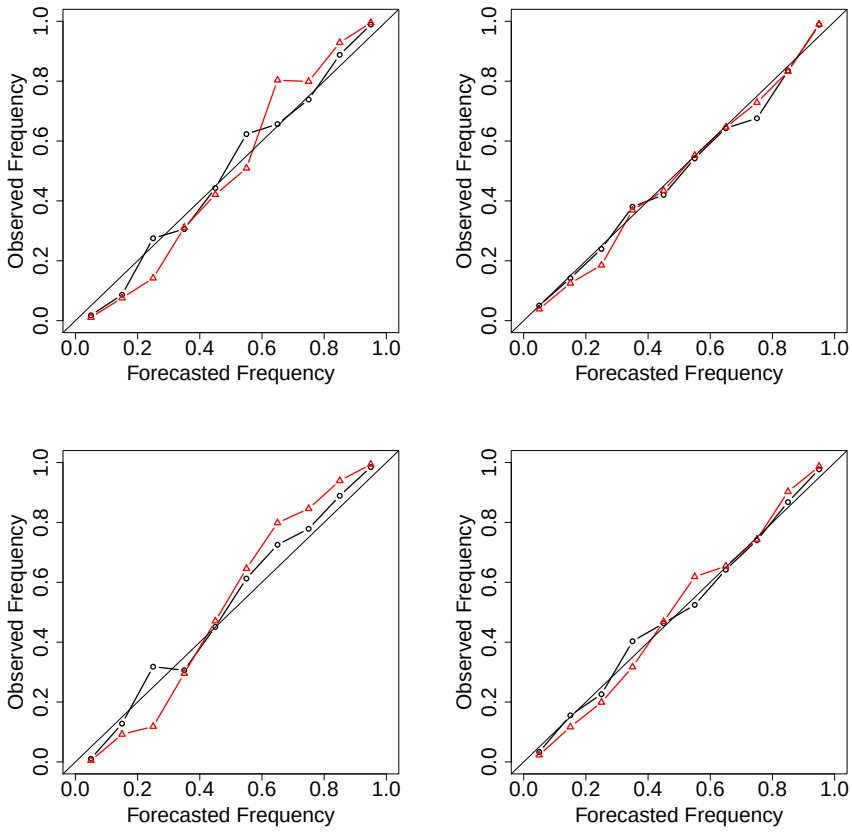


Figure 7.15 – Reliability diagrams of the raw ensemble (red) and weighted ensemble (black) for midday hours: 30-min lead time Réunion (top left), 3-h lead time Réunion (top right), 30-min lead time Corsica (bottom left), 3-h lead time Corsica (bottom right).

combined with an online learning algorithm to calibrate the weighted forecast, which outperformed almost always the individual forecasts. Consequently, we have shown that sequential aggregation enabled a seamless blending of the multiple sources of information for each lead time. For short lead times, the large improvement of the forecast quality was due to the rolling quantiles estimated with real-time production data. Satellite information proved to be very useful for Corsica, especially below 2 h of lead time. In contrast, we did not identify a gain brought by satellite information for Réunion.

Further work may focus on statistical modeling of satellite data. Indeed, the single statistical model used for all hours and lead times could be refined. Improved persistence forecast with ARMA (autoregressive moving average) models could also be of interest to quantify the gains brought by satellite data. Such study may include the identification of scenarios where satellite information demonstrates its value over time series analysis of production data. Furthermore, probabilistic forecasting of ramp events as already achieved for wind power in Bossavy et al. [BGK13] or in the review Gallego-Castillo et al. [GCL15] would give new indicators to grid operators, going beyond the confidence intervals of each half-hour of the day. Additional studies could also investigate other information sources such as AROME very short term forecasts (updated every hour) or sky cameras data.

Appendix 7.A Empirical results of quantile-weighted scoring rules with real-world data

Quantile weighted scoring rules, introduced in Chapter 4, offer an alternative to the CRPS. They allow higher focus on either the distribution tails or the distribution center. Examples of quantile-weighted scores are shown in Table 7.2, with the notation of Chapter 4. In this appendix, we provide empirical results on quantile-weighted scoring rules with real-world data. We focus on the data set of Réunion, because similar results were found for Corsica and for the data set mixing ECMWF, AROME and other Météo France forecasts of Chapter 6.

First, we investigate whether similar results are obtained with the quantile-weighted scoring rules described in Table 7.2. In the remaining of this thesis, the CRPS is the main score for probabilistic forecasts evaluation. Hence we wish to know whether the same conclusions may be drawn with other evaluation tools. Skill scores against

$\omega(\alpha)$	Score $S(\mathrm{G}, y)$	
$(\alpha(1-\alpha))^{-1}$	$-\int \mathrm{H}_y \ln \mathrm{G} + (1-\mathrm{H}_y)\ln(1-\mathrm{G})$	(CRIGN)
$(\alpha(1-\alpha))^{-1/2}$	$\int \mathrm{H}_y \arcsin(\sqrt{1-\mathrm{G}}) + (1-\mathrm{H}_y)\arcsin(\sqrt{\mathrm{G}}) - \sqrt{\mathrm{G}(1-\mathrm{G})}$	(arcsin-CRPS)
2	$\int (\mathrm{H}_y - \mathrm{G})^2$	(CRPS)
$\alpha(1-\alpha)$	$\int \mathrm{H}_y(\frac{(1-\mathrm{G})^3}{3} - \frac{(1-\mathrm{G})^4}{4}) + (1-\mathrm{H}_y)(\frac{\mathrm{G}^3}{3} - \frac{\mathrm{G}^4}{4})$	(cubic-CRPS)

Table 7.2 – Example of quantile-weighted scores $S(\mathrm{G}, y)$ with weighting functions ω. We plateau the log terms in the CRIGN to -log(0.001) (rounded to 6.908) to avoid infinite-valued scores.

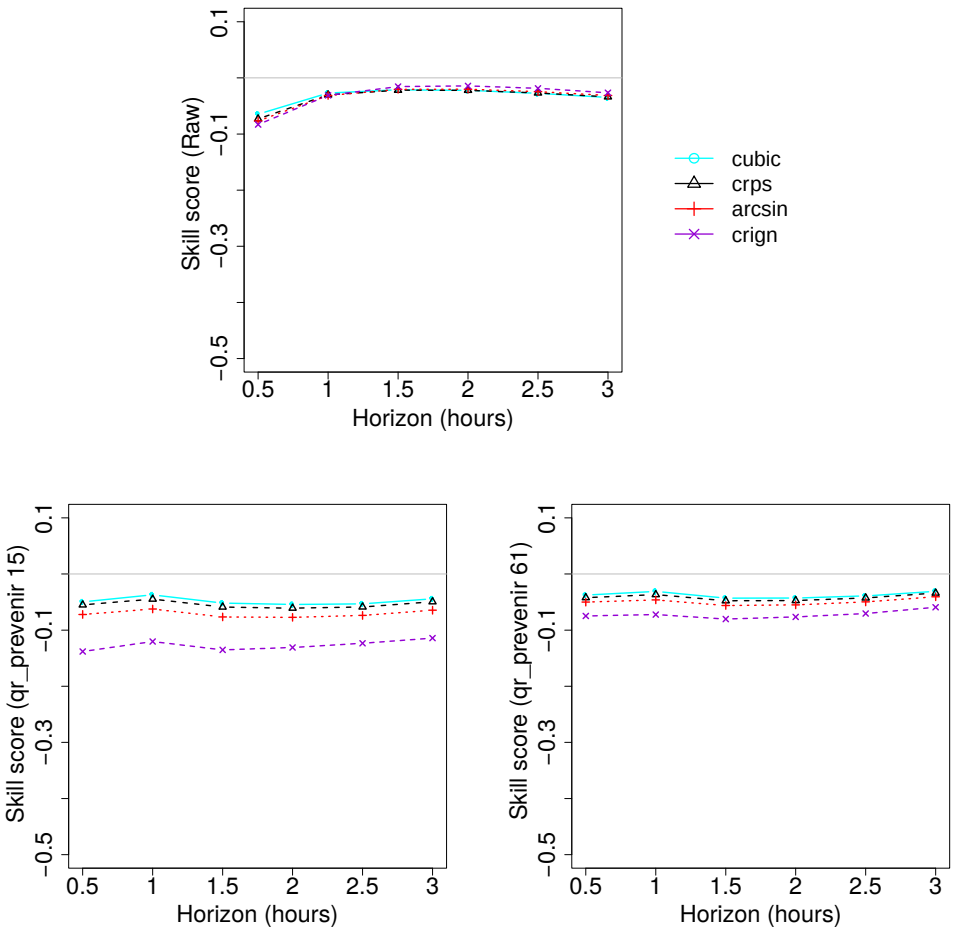


Figure 7.16 – Skill scores against our weighted forecast Pond for the cubic-CRPS, the CRPS, the arcsin-CRPS and the CRIGN. The subensembles Raw (top), "qr_prevenir" with 15 quantiles (bottom left) and "qr_prevenir" with 61 quantiles (bottom right) are evaluated.

our weighted forecasts Pond are shown for several quantile-weighted scoring rules in Figure 7.16 for the Réunion data set. Only little variation is noticed in the skill scores of the raw ensemble, whatever the chosen evaluation metrics. This results is an indicator of the robustness of the results obtained with the CRPS. In the skill scores of the subensemble "qr_prevenir", we find a difference up to 10% between the evaluation with the cubic-CRPS and with the CRIGN, indicating that the subensemble provides a better description of the distribution central quantiles than the distribution tails. A better discretization of the distribution tails of "qr_prevenir" mainly improves the CRIGN skill score. This is shown by increasing the subensemble size to 61 quantiles instead of the initial 15 quantiles. Increasing the subensemble size also decreases the difference in the skill scores between the CRIGN and the cubic-CRPS to 5%. As already mentioned, we did not find any advantage of running sequential aggregation algorithms with 61 rolling quantiles in each sub ensembles.

Evaluation		CRPS	
Optimization	CRPS	arcsin-CRPS	CRIGN
0.50 h	0.0213	0.0213	0.0214
1.00 h	0.0287	0.0288	0.0290
1.50 h	0.0332	0.0333	0.0337
2.00 h	0.0365	0.0368	0.0372
2.50 h	0.0390	0.0393	0.0397
3.00 h	0.0409	0.0410	0.0415
Evaluation		arcsin-CRPS	
Optimization	CRPS	arcsin-CRPS	CRIGN
0.50 h	0.0267	0.0266	0.0266
1.00 h	0.0355	0.0355	0.0357
1.50 h	0.0409	0.0410	0.0414
2.00 h	0.0450	0.0452	0.0457
2.50 h	0.0481	0.0483	0.0487
3.00 h	0.0504	0.0504	0.0510
Evaluation		CRIGN	
Optimization	CRPS	arcsin-CRPS	CRIGN
0.50 h	0.0733	0.0727	0.0726
1.00 h	0.0965	0.0962	0.0967
1.50 h	0.1115	0.1112	0.1122
2.00 h	0.1226	0.1226	0.1236
2.50 h	0.1313	0.1313	0.1322
3.00 h	0.1381	0.1374	0.1387

Table 7.3 – Daily averages of quantile-weighted scores according to the optimization criterion and the lead time.

Secondly, we run online learning experiment with gradients of quantile-weighted scores as loss function. We recall that these gradients are introduced in Section 4.1.2. We generate new weighted forecasts based on the optimization of the arcsin-CRPS and the CRIGN. The algorithm ML-Poly is run because it is parameter-free and its weight update scales well with the order of magnitude of the loss. For the CRIGN gradients, the terms $1/(\alpha(1-\alpha))$ are clipped between between 4 and 20. Consequently, quantiles of order below 5% and above 95% are equally weighted by the weighting function ω. Without surprise, we found differences in the weights of each member depending on the loss function, but without significant impact on the weighted forecast. Indeed, weighted forecasts learned with the CRPS, the arcsin-CRPS or the CRIGN show relative score differences below 1% in terms of CRPS, arcsin-CRPS and the CRIGN, as indicated in Table 7.3. Other evaluation tools such as the rank histogram or the reliability diagram did not help us to discriminate the weighted forecasts.

As a conclusion, the results of this Appendix are mainly robustness of the results in favor of the CRPS for both evaluation and learning. Indeed, the comparison of the raw ensemble against the weighted forecast (learned with the CRPS) indicates similar skill scores for various performance metrics. Besides, online learning with the CRPS, the arcsin-CRPS and the CRIGN provided very similar weighted distributions. Further study should investigate a wider range of data sets, concerning for example extreme-

value verification with quantile-weighted scores focusing on the distribution tails. Score sensitivity to correct quantile estimation is another direction for future research as well as the impact of the ensemble size.

8 Thesis conclusions

Our case studies demonstrate that combining forecasts derived from several meteorological centers or postprocessing techniques enables to improve the accuracy of solar radiation and PV power forecasts. To do so, we resorted to online learning techniques providing update rules of the combination weights. These techniques come with theoretical performance guarantees on the predictive power of the combination of the forecasts, under essentially no assumptions. These methods do not depend on the nature of the output variable, but their implementation on weather-related variables (solar radiation and PV power) are particularly interesting under at least two aspects: the large uncertainty of the forecasts due to the difficulty to forecast clouds at the correct time and location, and the spatial structure of the forecasts.

The following practical results are emphasized in our case studies:

— The weighted combination of forecasts more than often outperforms uniform averages of a subset of forecasts.

— Including forecasts of poor quality in the ensemble does not degrade the accuracy of the weighted forecasts.

— Online learning techniques minimizing the CRPS improve the calibration and the reliability of probabilistic forecasts, as verified for several evaluation tools (CRPS, rank histogram, spread skill diagram, and reliability diagram).

We now briefly review more specific results obtained for each case study. Chapter 2 focuses on non-probabilistic forecasting of solar radiation. Several TIGGE ensembles of solar forecasts are compared against satellite-derived HelioClim maps. This work is among the first examples of sequential aggregation techniques applied to maps of forecasts with the work of Baudin [Bau15] and Zamo [Zam16]. Our forecast favorably compares against HRES (ECMWF reference forecast) and spatial patterns are more finely described.

For PV power forecasting, statistical models between meteorological and production data are built, which brings an additional challenge. Both the meteorological and the conversion model inaccuracies should be taken into account when delivering probabilistic forecasts of PV power. Chapter 3 details a proof-of-concept case study with ECMWF and Météo France forecasts, where online learning techniques meet CRPS minimization and the calibration of probabilistic forecasts. This case study covers 219 PV sites, with the high 30-min temporal resolution. Even for the long lead time of 6 days, the superiority of weather forecasts over climatological averages is verified.

We include AROME forecasts in our study in Chapter 6. A large amount of members are generated to take into account the rich spatio-temporal information of AROME. Statistically calibrated forecasts from quantile regressions greatly improve the CRPS of the weighted forecasts. Besides, for short lead times below 24 h, the weighted combination of PV power forecasts generated with only AROME and HRES weather forecasts

shows almost the same performance as the weighted forecast including all available power forecasts.

The study of intraday updates for insular systems is undertaken in Chapter 7 for Réunion and Corsica, where intraday satellite-derived forecasts and day-to-day weather-derived forecasts are blended. The major impact of integrating the latest available information is demonstrated. Besides, satellite information proves to enhance the forecast accuracy for Corsica but not for Réunion.

Online learning methods face practical limitations due to data unavailability. While it is always possible to replace missing data with moving averages, other ideas for missing data imputations are described below. In case of missing PV power observations, the average of nearby power plants productions may be used as replacing value. If the forecaster is only interested in the total production, adjusting the total installed capacity is an easy-to-implement solution. In case of missing forecasts, it may be possible to run the online learning experiment since its beginning by including only the available experts. Besides, the setting of sleeping (or specialized) experts is well-studied, especially for the algorithm ML-Poly. In a few words, the weights are allowed to incorporate prior knowledge through the arbitrary confidences $I_{m,t}$, which are also integrated in the regret bound. The weight $u_{m,t}$ becomes proportional to $I_{m,t}u_{m,t}$, and the loss $\ell_{m,t}$ becomes $I_{m,t}\ell_{m,t} + (1 - I_{m,t}) \sum_{k \leqslant M} u_{k,t}\ell_{k,t}$.

In our case studies, we encountered several open questions that we summarize below and leave for future research. Several of them concern the multidimensionality of the output variable.

1. In Chapter 2, it is shown that sequential aggregation improves solar forecasts. A new expert for PV forecasts can therefore be built using improved solar forecasts. What is the gain in PV power performance that is achievable thanks to this new expert, and for which spatio-temporal scales?
2. Can statistical errors from the conversion model and meteorological errors be separated? A first step would be to build a perfect statistical model with weather observations or satellite observations. However, observational noise and representativeness errors prevent from building this perfect model.
3. Better solar forecasts are obtained by running sequential aggregations independently at each grid point. Could spatial patterns be included in this setting? An idea would be to apply sequential aggregation over spatial averages covering specific regions. The definition of such spatial subsets and their evolution over time is however unclear.
4. We focused on the marginal distribution of each time step. Can we extend our framework to include forecast trajectories and probability of ramp events as in Bossavy et al. [BGK13] or in the review Gallego-Castillo et al. [GCL15] for wind power? This may require multivariate experts in the time dimension.
5. The CRPS generalizes to the energy score for multivariate probabilistic forecasts [Sze03; GR07]. Probabilistic multivariate forecasting therefore seems compatible with online learning, provided that multivariate experts are available. We showed that quantile forecasts greatly improve the quality of the forecasts. Could multivariate quantile forecasts be generated? Possible applications include the

challenging task of joint forecasting wind, PV and the electricity demand.

6. Say expert k performs significantly better than expert m in situation A_k, and conversely say expert k performs significantly worst than expert m in situation A_m. How should this prior knowledge be integrated? The forecaster may give prior weights to the experts m and k according to its opinion that situation A_k or A_m will occur, or the forecaster may build a new expert that is a combination of the experts m and k. This new expert should be closer to expert k in situation A_k and closer to expert m in situation A_m. Our opinion is that the latter setting allows for a greater flexibility.

Non-local strictly proper scoring rules are investigated in Chapter 3 for the CRPS and more generally in Chapter 4. Explanations for the CRPS bias and the definition of a fair CRPS with classes of members are proposed. Besides, we introduced improved formulations for threshold-weighted and quantile-weighted scoring rules, as well as better understanding of these scores for model mixtures. In the case study of Appendix 7.A, results obtained with the CRPS or other scoring rules are not significantly different. This is a robustness argument in favor of the CRPS. For the verification of extreme events, it might be of interest to compare strongly asymmetric quantile-weighted scoring rules and simple quantile scores.

We defined a generalized least-square alternative to the CRPS in Section 4.2, which includes the observational distribution. New perspectives are brought on the connections between Pearson's χ^2, the Anderson-Darling test, and Cramer-von Mises test. This work appears as a starting point to further theoretical analysis. Case studies are needed to compare this new loss against common learning and verification procedures.

Bibliography

[Ait36] Aitken, A. C. « On Least Squares and Linear Combination of Observations ». In: *Proceedings of the Royal Society of Edinburgh* 55 (1936), pp. 42–48.

[Ale+15] Alessandrini, S. et al. « An analog ensemble for short-term probabilistic solar power forecast ». In: *Applied Energy* 157 (2015), pp. 95–110.

[APN15] Almeida, M. P. et al. « PV power forecast using a nonparametric PV model ». In: *Solar Energy* 115 (2015), pp. 354–368. ISSN: 0038-092X.

[And96] Anderson, J. L. « A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations ». In: *Journal of Climate* 9.7 (1996), pp. 1518–1530.

[Ant+16] Antonanzas, J. et al. « Review of photovoltaic power forecasting ». In: *Solar Energy* 136 (2016), pp. 78–111.

[Atg04] Atger, F. « Estimation of the reliability of ensemble-based probabilistic forecasts ». In: *Quarterly Journal of the Royal Meteorological Society* 130.597 (2004), pp. 627–646.

[Aud09] Audibert, J.-Y. « Fast learning rates in statistical inference through aggregation ». In: *The Annals of Statistics* 37.4 (2009), pp. 1591–1646.

[AW01] Azoury, K. S. and Warmuth, M. K. « Relative loss bounds for on-line density estimation with the exponential family of distributions ». In: *Machine Learning* 43.3 (2001), pp. 211–246.

[BMN09] Bacher, P. et al. « Online short-term solar power forecasting ». In: *Solar Energy* 83.10 (2009), pp. 1772–1783.

[Bad+15] Badosa, J. et al. « Reliability of day-ahead solar irradiance forecasts on Reunion Island depending on synoptic wind and humidity conditions ». In: *Solar Energy* 115 (2015), pp. 306–321. ISSN: 0038-092X.

[Bau15] Baudin, P. « Prévision séquentielle par agrégation d'ensemble: application à des prévisions météorologiques assorties d'incertitudes ». PhD thesis. Paris-Sud XI, 2015.

[BPS09] Ben-David, S. et al. « Agnostic Online Learning. » In: *COLT*. 2009.

[Ben10] Benedetti, R. « Scoring rules for forecast verification ». In: *Monthly Weather Review* 138.1 (2010), pp. 203–211.

[BF14] Bentzien, S. and Friederichs, P. « Decomposition and graphical portrayal of the quantile score ». In: *Quarterly Journal of the Royal Meteorological Society* 140.683 (2014), pp. 1924–1934.

[Ber79] Bernardo, J. M. « Expected information as expected utility ». In: *The Annals of Statistics* (1979), pp. 686–690.

[BP11] Biau, G. and Patra, B. « Sequential quantile prediction of time series ». In: *Information Theory, IEEE Transactions on* 57.3 (2011), pp. 1664–1674.

[Bla+11] Blanc, P. et al. « The HelioClim Project: Surface Solar Irradiance Data for Climate Applications ». In: *Remote Sensing* 3.2 (2011), pp. 343–361. ISSN: 2072-4292.

[BGK13] Bossavy, A. et al. « Forecasting ramps of wind power production with numerical weather prediction ensembles ». In: *Wind Energy* 16.1 (2013), pp. 51–63.

[Bou+10] Bougeault, P. et al. « The THORPEX interactive grand global ensemble ». In: *Bulletin of the American Meteorological Society* 91.8 (2010), pp. 1059–1072.

[Bow06] Bowler, N. E. « Explicitly accounting for observation error in categorical verification of forecasts ». In: *Monthly weather review* 134.6 (2006), pp. 1600–1606.

[Bow08] Bowler, N. E. « Accounting for the effect of observation errors on verification of MOGREPS ». In: *Meteorological Applications* 15.1 (2008), pp. 199–205. ISSN: 1469-8080.

[Bre01] Breiman, L. « Random forests ». In: *Machine learning* 45.1 (2001), pp. 5–32.

[Bri50] Brier, G. W. « Verification of Forecasts Expressed in Terms of Probability ». In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[Brö09] Bröcker, J. « Reliability, sufficiency, and the decomposition of proper scores ». In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (2009), pp. 1512–1519.

[Brö12] Bröcker, J. « Evaluating raw ensembles with the continuous ranked probability score ». In: *Quarterly Journal of the Royal Meteorological Society* 138.667 (2012), pp. 1611–1617.

[BS07a] Bröcker, J. and Smith, L. A. « Increasing the reliability of reliability diagrams ». In: *Weather and forecasting* 22.3 (2007), pp. 651–661.

[BS07b] Bröcker, J. and Smith, L. A. « Scoring probabilistic forecasts: The importance of being proper ». In: *Weather and Forecasting* 22.2 (2007), pp. 382–388.

[Bui+05] Buizza, R. et al. « A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems ». In: *Monthly Weather Review* 133.5 (2005), pp. 1076–1097.

[BSS05] Buja, A. et al. « Loss functions for binary class probability estimation and classification: Structure and applications ». In: *Working draft* (2005). URL: stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf.

[CT05] Candille, G. and Talagrand, O. « Evaluation of probabilistic prediction systems for a scalar variable ». In: *Quarterly Journal of the Royal Meteorological Society* 131.609 (2005), pp. 2131–2150. ISSN: 1477-870X.

[CT08] Candille, G. and Talagrand, O. « Impact of observational error on the validation of ensemble prediction systems ». In: *Quarterly Journal of the Royal Meteorological Society* 134.633 (2008), pp. 959–971. ISSN: 1477-870X.

[Cat04] Catoni, O. « Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, Ecole d'été de Probabilités de Saint-Flour XXXI–2001 ». In: *Lecture Notes in Mathematics* 1851 (2004), pp. 1–269.

[CL03] Cesa-Bianchi, N. and Lugosi, G. « Potential-based algorithms in on-line prediction and game theory ». In: *Machine Learning* 51.3 (2003), pp. 239–261.

[CL06] Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

[CMS05] Cesa-Bianchi, N. et al. « Improved second-order bounds for prediction with expert advice ». In: *International Conference on Computational Learning Theory*. Springer. 2005, pp. 217–232.

[CSS11] Cesa-Bianchi, N. et al. « Online learning of noisy data ». In: *IEEE Transactions on Information Theory* 57.12 (2011), pp. 7907–7931.

[CW99] Clemen, R. T. and Winkler, R. L. « Combining probability distributions from experts in risk analysis ». In: *Risk analysis* 19.2 (1999), pp. 187–203.

[Cou+91] Courtier, P. et al. « The Arpège project at Météo-France ». In: *ECMWF Seminar Proceedings*. Vol. 2. 1991, pp. 193–231.

[Dam+14] Dambreville, R. et al. « Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model ». In: *Renewable Energy* 72 (2014), pp. 291–300.

[Daw08] Dawid, A. « Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds ». In: *TEST* 17.2 (2008), pp. 243–244. ISSN: 1133-0686. DOI: 10.1007/s11749-008-0118-6. URL: http://dx.doi.org/10.1007/s11749-008-0118-6.

[Deh+14] Dehghan, A. et al. « Evaluation and improvement of TAPM in estimating solar irradiance in Eastern Australia ». In: *Solar Energy* 107 (2014), pp. 668–680. ISSN: 0038-092X.

[Des+15] Descamps, L. et al. « PEARP, the Météo-France short-range ensemble prediction system ». In: *Quarterly Journal of the Royal Meteorological Society* 141.690 (2015), pp. 1671–1685. ISSN: 1477-870X.

[Dev+13] Devaine, M. et al. « Forecasting electricity consumption by aggregating specialized experts ». In: *Machine Learning* 90.2 (2013), pp. 231–260. ISSN: 0885-6125.

[Ehm+16] Ehm, W. et al. « Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings ». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.3 (2016), pp. 505–562.

[Elm05] Elmore, K. L. « Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts ». In: *Weather and forecasting* 20.5 (2005), pp. 789–795.

[Eps69] Epstein, E. S. « A Scoring System for Probability Forecasts of Ranked Categories ». In: *Journal of Applied Meteorology and Climatology* 8.6 (1969), pp. 985–987.

[Esp+10] Espinar, B. et al. « Photovoltaic Forecasting: A state of the art ». In: *Proceedings 5th European PV-Hybrid and Mini-Grid Conference*. 2010.

[Fer14] Ferro, C. « Fair scores for ensemble forecasts ». In: *Quarterly Journal of the Royal Meteorological Society* 140.683 (2014), pp. 1917–1923.

[FRW08] Ferro, C. A. T. et al. « On the effect of ensemble size on the discrete and continuous ranked probability scores ». In: *Meteorological Applications* 15.1 (2008), pp. 19–24. ISSN: 1469-8080. DOI: 10.1002/met.45.

[For+14] Fortin, V. et al. « Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? » In: *Journal of Hydrometeorology* 15.4 (2014), pp. 1708–1713.

[FRG10] Fraley, C. et al. « Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging ». In: *Monthly Weather Review* 138.1 (2010), pp. 190–202.

[FV14] Frénay, B. and Verleysen, M. « Classification in the presence of label noise: a survey ». In: *IEEE transactions on neural networks and learning systems* 25.5 (2014), pp. 845–869.

[FFS13] Fricker, T. E. et al. « Three recommendations for evaluating climate predictions ». In: *Meteorological Applications* 20.2 (2013), pp. 246–255. ISSN: 1469-8080. DOI: 10.1002/met.1409. URL: http://dx.doi.org/10.1002/met.1409.

[GSE14] Gaillard, P. et al. « A Second-order Bound with Excess Losses ». In: *Proceedings of COLT'14*. Vol. 35. JMLR: Workshop and Conference Proceedings, 2014, pp. 176–196.

[GGN16] Gaillard, P. et al. « Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting ». In: *International Journal of Forecasting* 32.3 (2016), pp. 1038–1050.

[GCL15] Gallego-Castillo, C. et al. « A review on the recent history of wind power ramp forecasting ». In: *Renewable and Sustainable Energy Reviews* 52 (2015), pp. 1148–1157.

[GDM80] Gautier, C. et al. « A simple physical model to estimate incident solar radiation at the surface from GOES satellite data. » In: *Journal of Applied Meteorology* 19 (Aug. 1980), pp. 1005–1012.

[GM90] Genest, C. and McConway, K. J. « Allocating the weights in the linear opinion pool ». In: *Journal of Forecasting* 9.1 (1990), pp. 53–73. ISSN: 1099-131X. DOI: 10.1002/for.3980090106. URL: http://dx.doi.org/10.1002/for.3980090106.

[Ger11] Gerchinovitz, S. « Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques ». PhD thesis. Université Paris Sud-Paris XI, 2011.

[GA11] Geweke, J. and Amisano, G. « Optimal prediction pools ». In: *Journal of Econometrics* 164.1 (2011), pp. 130–141.

[GK14] Gneiting, T. and Katzfuss, M. « Probabilistic forecasting ». In: *Annual Review of Statistics and Its Application* 1 (2014), pp. 125–151.

[GR07] Gneiting, T. and Raftery, A. E. « Strictly proper scoring rules, prediction, and estimation ». In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[GR11] Gneiting, T. and Ranjan, R. « Comparing density forecasts using threshold- and quantile-weighted scoring rules ». In: *Journal of Business & Economic Statistics* 29.3 (2011).

[Gne+05] Gneiting, T. et al. « Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation ». In: *Monthly Weather Review* 133.5 (2005), pp. 1098–1118.

[Gon+10] González Abril, L. et al. « The similarity between the square of the coefficient of variation and the Gini index of a general random variable ». In: *Revista de métodos cuantitativos para la economía y la empresa* 10 (2010), pp. 5–18.

[Goo52] Good, I. J. « Rational decisions ». In: *Journal of the Royal Statistical Society. Series B (Methodological)* 14 (1952), pp. 107–114.

[Gri+06] Grimit, E. P. et al. « The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification ». In: *Quarterly Journal of the Royal Meteorological Society* 132 (2006), pp. 2925–2942.

[Hag+12] Hagedorn, R. et al. « Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts ». In: *Quarterly Journal of the Royal Meteorological Society* 138.668 (2012), pp. 1814–1827. ISSN: 1477-870X.

[Ham01] Hamill, T. M. « Interpretation of rank histograms for verifying ensemble forecasts ». In: *Monthly Weather Review* 129.3 (2001), pp. 550–560.

[HC97] Hamill, T. M. and Colucci, S. J. « Verification of Eta/RSM Short-Range Ensemble Forecasts ». In: *Monthly Weather Review* 125 (1997), pp. 1312–1327.

[Ham+99] Hammer, A. et al. « Short-term forecasting of solar radiation: a statistical approach using satellite data ». In: *Solar Energy* 67.1–3 (1999), pp. 139–150.

[HAK07] Hazan, E. et al. « Logarithmic regret algorithms for online convex optimization ». In: *Machine Learning* 69.2-3 (2007), pp. 169–192.

[Her00] Hersbach, H. « Decomposition of the continuous ranked probability score for ensemble prediction systems ». In: *Weather and Forecasting* 15.5 (2000), pp. 559–570.

[Hof+13] Hoff, T. E. et al. « Reporting of irradiance modeling relative prediction errors ». In: *Progress in Photovoltaics: Research and Applications* 21.7 (2013), pp. 1514–1519.

[HP15] Huang, J. and Perry, M. « A semi-empirical approach using gradient boosting and -nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting ». In: *International Journal of Forecasting* (2015), ISSN: 0169-2070.

[IPC13] Inman, R. H. et al. « Solar forecasting methods for renewable energy integration ». In: *Progress in Energy and Combustion Science* 39.6 (2013), pp. 535–576.

[JS12] Jolliffe, I. and Stephenson, D. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 2012. ISBN: 9781119961079.

[JMA15] Junk, C. et al. « Analog-Based Ensemble Model Output Statistics ». In: *Monthly Weather Review* 143.7 (2015), pp. 2909–2917. DOI: 10.1175/MWR-D-15-0095.1.

[Kal60] Kalman, R. E. « A new approach to linear filtering and prediction problems ». In: *Journal of basic Engineering* 82.1 (1960), pp. 35–45.

[KW97] Kivinen, J. and Warmuth, M. K. « Exponentiated Gradient versus Gradient Descent for Linear Predictors ». In: *Information and Computation* 132.1 (1997), pp. 1–63. ISSN: 0890-5401. DOI: http://dx.doi.org/10.1006/inco.1996.2612.

[KB78] Koenker, R. and Bassett Jr, G. « Regression quantiles ». In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.

[KH01] Koenker, R. and Hallock, K. « Quantile regression: An introduction ». In: *Journal of Economic Perspectives* 15.4 (2001), pp. 43–56.

[KV15] Koolen, W. M. and Van Erven, T. « Second-order Quantile Methods for Experts and Combinatorial Games. » In: *COLT*. Vol. 40. 2015, pp. 1155–1175.

[LT07] Laio, F. and Tamea, S. « Verification tools for probabilistic forecasts of continuous hydrological variables ». In: *Hydrology and Earth System Sciences Discussions* 11.4 (2007), pp. 1267–1277.

[LLD16] Lauret, P. et al. « Solar Forecasting in a Challenging Insular Context ». In: *Atmosphere* 7.2 (2016). ISSN: 2073-4433.

[Ler+15] Lerch, S. et al. « Forecaster's dilemma: Extreme events and forecast evaluation ». In: *arXiv preprint arXiv:1512.09244* (2015).

[LP08] Leutbecher, M. and Palmer, T. N. « Ensemble forecasting ». In: *Journal of Computational Physics* 227.7 (2008), pp. 3515–3539.

[Lew05] Lewis, J. M. « Roots of ensemble forecasting ». In: *Monthly weather review* 133.7 (2005), pp. 1865–1885.

[LW94] Littlestone, N. and Warmuth, M. K. « The Weighted Majority Algorithm ». In: *Inf. Comput.* 108.2 (Feb. 1994), pp. 212–261. ISSN: 0890-5401.

[Lor+09a] Lorenz, E. et al. « Benchmarking of different approaches to forecast solar irradiance ». In: *Proceedings of the 24th European Photovoltaic Solar Energy Conference*. 2009, pp. 4199–4208.

[Lor+09b] Lorenz, E. et al. « Irradiance forecasting for the power prediction of grid-connected photovoltaic systems ». In: *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* 2.1 (2009), pp. 2–10.

[LKH12] Lorenz, E. et al. « Short term forecasting of solar irradiance by combining satellite data and numerical weather predictions ». In: *Proceedings of the 27th European PV Solar Energy Conference (EU PVSEC), Frankfurt, Germany*. Vol. 2428. 2012, p. 44014405.

[LS15] Luo, H. and Schapire, R. E. « Achieving All with No Parameters: AdaNormalHedge ». In: *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*. 2015, pp. 1286–1304.

[Mal10] Mallet, V. « Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation ». In: *Journal of Geophysical Research* 115.D24303 (2010).

[MMS07] Mallet, V. et al. *Description of Sequential Aggregation Methods and their Performances for Ozone Ensemble Forecasting*. Tech. rep. DMA-07-08. École normale supérieure de Paris, 2007.

[MSM09] Mallet, V. et al. « Ozone ensemble forecast with machine learning algorithms ». In: *Journal of Geophysical Research* 114.D05307 (2009).

[MNZ13] Mallet, V. et al. « Minimax filtering for sequential aggregation: Application to ensemble forecast of ozone analyses ». In: *Journal of Geophysical Research* 118.11 (2013), pp. 11, 294–11, 303.

[MW76] Matheson, J. E. and Winkler, R. L. « Scoring rules for continuous probability distributions ». In: *Management science* 22.10 (1976), pp. 1087–1096.

[McM11] McMahan, H. B. « Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization. » In: *AISTATS*. 2011, pp. 525–533.

[Mei06] Meinshausen, N. « Quantile regression forests ». In: *Journal of Machine Learning Research* 7.Jun (2006), pp. 983–999.

[MS13] Merkle, E. C. and Steyvers, M. « Choosing a strictly proper scoring rule ». In: *Decision Analysis* 10.4 (2013), pp. 292–304.

[Mor91] Morcrette, J.-J. « Radiation and cloud radiative properties in the European Centre for Medium Range Weather Forecasts forecasting system ». In: *Journal of Geophysical Research: Atmospheres* 96.D5 (1991), pp. 9121–9132. ISSN: 2156-2202.

[Mur71] Murphy, A. H. « A Note on the Ranked Probability Score ». In: *Journal of Applied Meteorology and Climatology* 10.2 (1971), pp. 155–156.

[Mur73] Murphy, A. H. « A New Vector Partition of the Probability Score ». In: *Journal of Applied Meteorology and Climatology* 12.4 (1973), pp. 595–600.

[Nat+13] Natarajan, N. et al. « Learning with noisy labels ». In: *Advances in neural information processing systems*. 2013, pp. 1196–1204.

[NMN06] Nielsen, H. A. et al. « Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts ». In: *Wind Energy* 9.1-2 (2006), pp. 95–108. ISSN: 1099-1824.

[Ohm+98] Ohmura, A. et al. « Baseline Surface Radiation Network (BSRN/WCRP): New precision radiometry for climate research ». In: *Bulletin of the American Meteorological Society* 79.10 (1998), pp. 2115–2136.

[OCC15] Orabona, F. et al. « A generalized online mirror descent with applications to classification and regression ». In: *Machine Learning* 99.3 (2015), pp. 411–435.

[Pal+09] Palmer, T. et al. *Stochastic parametrization and model uncertainty*. European Centre for Medium-Range Weather Forecasts, 2009.

[Per+13] Perez, R. et al. « Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe ». In: *Solar Energy* 94 (2013), pp. 305–326.

[PSZ97] Perez, R. et al. « Comparing satellite remote sensing and ground network measurements for the production of site/time specific irradiance data ». In: *Solar Energy* 60.2 (1997), pp. 89–96.

[Raf+05] Raftery, A. E. et al. « Using Bayesian model averaging to calibrate forecast ensembles ». In: *Monthly Weather Review* 133 (2005), pp. 1, 155–1, 174.

[RG10] Ranjan, R. and Gneiting, T. « Combining probability forecasts ». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.1 (2010), pp. 71–91.

[Rao02] Rao, C. « Karl Pearson chi-square test the dawn of statistical inference ». In: *Goodness-of-fit tests and model validity*. Springer, 2002, pp. 9–24.

[RSS15] Ren, Y. et al. « Ensemble methods for wind and solar power forecasting—A state-of-the-art review ». In: *Renewable and Sustainable Energy Reviews* 50 (2015), pp. 82–91. ISSN: 1364-0321.

[RLW04] Rigollier, C. et al. « The method Heliosat-2 for deriving shortwave solar radiation from satellite images ». In: *Solar Energy* 77.2 (2004), pp. 159–169. ISSN: 0038-092X.

[Sae+04] Saetra, Ø. et al. « Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability ». In: *Monthly Weather Review* 132.6 (2004), pp. 1487–1501.

[Sav71] Savage, L. J. « Elicitation of personal probabilities and expectations ». In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801.

[Sch89] Schervish, M. J. « A General Method for Comparing Probability Assessors ». In: *astat* 17.4 (Dec. 1989), pp. 1856–1879.

[Sch14] Scheuerer, M. « Probabilistic quantitative precipitation forecasting using ensemble model output statistics ». In: *Quarterly Journal of the Royal Meteorological Society* 140.680 (2014), pp. 1086–1096.

[Sei+11] Seity, Y. et al. « The AROME-France Convective-Scale Operational Model ». In: *Monthly Weather Review* 139.3 (2011), pp. 976–991.

[Sha11] Shalev-Shwartz, S. « Online learning and online convex optimization ». In: *Foundations and Trends in Machine Learning* 4.2 (2011), pp. 107–194.

[SAE66] Shuford, E. H. et al. « Admissible probability measurement procedures ». In: *Psychometrika* 31.2 (1966), pp. 125–145. ISSN: 1860-0980.

[Slo+07] Sloughter, J. M. L. et al. « Probabilistic quantitative precipitation forecasting using Bayesian model averaging ». In: *Monthly Weather Review* 135.9 (2007), pp. 3209–3220.

[SGR10] Sloughter, J. M. et al. « Probabilistic wind speed forecasting using ensembles and Bayesian model averaging ». In: *Journal of the American Statistical Association* 105.489 (2010), pp. 25–35.

[Spe+15] Sperati, S. et al. « The "Weather Intelligence for Renewable Energies" Benchmarking Exercise on Short-Term Forecasting of Wind and Solar Power Generation ». In: *Energies* 8.9 (2015), pp. 9594–9619.

[SAM16] Sperati, S. et al. « An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting ». In: *Solar Energy* 133 (2016), pp. 437–450.

[Sto10] Stoltz, G. « Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique ». In: *Journal de la Société Française de Statistique* 151.2 (2010), pp. 66–106.

[Sze03] Szekely, G. J. « E-Statistics: The energy of statistical samples ». In: *Bowling Green State University, Department of Mathematics and Statistics Technical Report* 03-05 (2003), pp. 2000–2003.

[TVS99] Talagrand, O. et al. *Evaluation of Probabilistic Prediction System*. Proceedings of the ECMWF Workshop on Predictability. Reading, United Kingdom, 1999.

[TE13] Thelen, J.-C. and Edwards, J. M. « Short-wave radiances: comparison between SEVIRI and the Unified Model ». In: *Quarterly Journal of the Royal Meteorological Society* 139.675 (2013), pp. 1665–1679.

[TG10] Thorarinsdottir, T. L. and Gneiting, T. « Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression ». In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.2 (2010), pp. 371–388.

[Tho+15] Thorey, J. et al. « Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database ». In: *Solar Energy* 120 (Oct. 2015), pp. 232–243.

[TMB16] Thorey, J. et al. « Online learning with the CRPS for ensemble forecasting ». In: *Quarterly Journal of the Royal Meteorological Society* (2016).

[TA12] Tödter, J. and Ahrens, B. « Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition ». In: *Monthly Weather Review* 140.6 (2012), pp. 2005–2017.

[V+13] Vernay, C. et al. « Review of satellite-based surface solar irradiation databases for the engineering, the financing and the operating of photovoltaic systems ». In: *ISES Solar World Congress*. 2013.

[VZ09] Vovk, V. and Zhdanov, F. « Prediction With Expert Advice For The Brier Game ». In: *J. Mach. Learn. Res.* 10 (Dec. 2009), pp. 2445–2471. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1577069.1755868.

[Vov01] Vovk, V. « Competitive on-line statistics ». In: *International Statistical Review/Revue Internationale de Statistique* (2001), pp. 213–248.

[Wil09] Wilks, D. S. « Extending logistic regression to provide full-probability-distribution MOS forecasts ». In: *Meteorological Applications* 16.3 (2009), pp. 361–368. ISSN: 1469-8080.

[WM68] Winkler, R. L. and Murphy, A. H. « "Good" probability assessors ». In: *Journal of applied Meteorology* 7.5 (1968), pp. 751–758.

[Win17] Wintenberger, O. « Optimal learning with Bernstein online aggregation ». In: *Machine Learning* 106.1 (2017), pp. 119–141.

[Yan04] Yang, Y. « Combining forecasting procedures: some theoretical results ». In: *Econometric Theory* 20.01 (2004), pp. 176–222.

[YS12] Yitzhaki, S. and Schechtman, E. *The Gini Methodology: A primer on a statistical methodology*. Vol. 272. Springer Science & Business Media, 2012.

[Yok+12] Yokohata, T. et al. « Reliability of multi-model and structurally different single-model ensembles ». In: *Climate Dynamics* 39.3-4 (2012), pp. 599–616. ISSN: 0930-7575.

[Zam16] Zamo, M. « Post-traitements statistiques de prévisions de vent déterministes et d'ensembles sur une grille ». PhD thesis. Laboratoire Mathématiques et Informatique Appliquées (Paris), 2016.

[Zam+14] Zamo, M. et al. « A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production ». In: *Solar Energy* 105 (2014), pp. 804–816.